



The effects of task complexity manipulated by intentional reasoning demands on second language learners' speech performance: interaction with language proficiency and working memory

Anas Ahmad Awwad

Thesis submitted for the degree of
Doctor of Philosophy
in Applied Linguistics

School of Literature and Languages
Department of English Language and Applied Linguistics
University of Reading

September 2017

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Anas Ahmad Awwad

14th September 2017

Table of Contents

| | |
|---|------|
| List of Tables | xi |
| List of Figures | xiii |
| List of Appendices | xiv |
| List of Abbreviations | xv |
| Abstract | xvii |
| Acknowledgements..... | xix |
| 1 Chapter one: INTRODUCTION..... | 1 |
| 1.1 Rationale for the study | 1 |
| 1.2 Organisation of the study | 5 |
| 2 Chapter 2: LITERATURE REVIEW | 8 |
| 2.1 Introduction | 8 |
| 2.2 Definitions of Task..... | 8 |
| 2.3 Task-Based Language Teaching (TBLT)..... | 11 |
| 2.3.1 Definitions of TBLT | 12 |
| 2.3.2 Emergence of TBLT | 13 |
| 2.3.3 Future of TBLT..... | 14 |
| 2.3.4 A Cognitive approach to TBLT | 15 |
| 2.4 Models of speech production | 17 |
| 2.4.1 Levelt's (1989) monolingual speech model..... | 17 |
| 2.4.2 Kormos's (2006) bilingual speech model..... | 20 |

| | | |
|--------|--|----|
| 2.5 | Aspects of speech performance..... | 21 |
| 2.6 | Task Complexity | 24 |
| 2.6.1 | The Limited Attentional Capacity | 27 |
| 2.6.2 | The Cognition Hypothesis | 29 |
| 2.6.3 | Skehan’s LAC versus Robinson’s CH..... | 34 |
| 2.7 | Research on Task Complexity | 36 |
| 2.7.1 | Research on resource-dispersing variables | 36 |
| 2.7.2 | Research on resource-directing variables | 38 |
| 2.7.3 | Research on reasoning as a TC variable | 39 |
| 2.8 | Intentional reasoning | 42 |
| 2.9 | Individual differences and language performance | 48 |
| 2.10 | Language Proficiency | 49 |
| 2.10.1 | Measuring language proficiency..... | 50 |
| 2.10.2 | Language proficiency and language performance | 51 |
| 2.10.3 | Language proficiency and Task Complexity research..... | 51 |
| 2.11 | Working Memory | 53 |
| 2.11.1 | Baddeley’s model of working memory | 54 |
| 2.11.2 | Measuring working memory..... | 57 |
| 2.11.3 | Working memory and language performance | 58 |
| 2.11.4 | Working memory and TBLT research..... | 59 |
| 2.12 | Conclusion..... | 62 |

| | | |
|--------|---|----|
| 3 | Chapter 3: METHODOLOGY: STUDY ONE..... | 65 |
| 3.1 | Introduction..... | 65 |
| 3.2 | Aims of the study..... | 65 |
| 3.3 | Research Questions (RQ) & Hypotheses (H)..... | 65 |
| 3.4 | Study Design..... | 67 |
| 3.5 | Participants..... | 68 |
| 3.6 | Tasks and Materials..... | 69 |
| 3.7 | Pilot study..... | 73 |
| 3.8 | Ethical Procedures..... | 74 |
| 3.9 | Data Collection..... | 74 |
| 3.10 | Data Coding..... | 75 |
| 3.10.1 | Measures of analysis..... | 75 |
| 3.10.2 | Measures of complexity..... | 76 |
| 3.10.3 | Measures of accuracy..... | 77 |
| 3.10.4 | Measures of fluency..... | 78 |
| 3.10.5 | Inter-rate reliability..... | 80 |
| 3.11 | Data Analysis..... | 81 |
| 3.12 | Conclusion..... | 81 |
| 4 | Chapter 4: RESULTS: STUDY ONE..... | 82 |
| 4.1 | Introduction..... | 82 |
| 4.2 | Descriptive analysis..... | 82 |

| | | |
|-------|--|-----|
| 4.3 | Multivariate analysis of variance (MANOVA)..... | 83 |
| 4.4 | Effects of reasoning demands on oral performance | 85 |
| 4.4.1 | Effects of IR demands on syntactic complexity | 87 |
| 4.4.2 | Effects of IR demands on lexical complexity | 87 |
| 4.4.3 | Effects of IR demands on accuracy | 88 |
| 4.4.4 | Effects of IR demands on fluency..... | 89 |
| 4.5 | Effects of IR on perceptions of TD | 89 |
| 4.6 | Summary of key findings (Study One) | 92 |
| 5 | Chapter 5: DISCUSSION: STUDY ONE | 93 |
| 5.1 | Introduction | 93 |
| 5.2 | Overview of Key findings | 93 |
| 5.3 | Intentional reasoning and syntactic complexity | 94 |
| 5.4 | Intentional reasoning and lexical complexity | 95 |
| 5.5 | Intentional reasoning and accuracy | 98 |
| 5.6 | Intentional reasoning and fluency | 99 |
| 5.7 | Intentional reasoning and Task Difficulty..... | 100 |
| 5.8 | Intentional reasoning and CALF measures | 102 |
| 5.9 | Study One closing remarks | 105 |
| 5.10 | Study Two Proposal | 107 |
| 6 | Chapter 6: METHODOLOGY: STUDY TWO | 109 |
| 6.1 | Introduction | 109 |

| | | |
|-------|---|-----|
| 6.2 | Aims of study | 109 |
| 6.3 | Research Questions (RQ) & Hypotheses (H)..... | 110 |
| 6.4 | Study design | 111 |
| 6.5 | Participants..... | 112 |
| 6.6 | Tasks and instruments | 113 |
| 6.6.1 | The video tasks | 113 |
| 6.6.2 | Task instructions | 113 |
| 6.6.3 | Language Proficiency Tests..... | 114 |
| 6.6.4 | Working Memory Tests | 116 |
| 6.6.5 | Task Difficulty Questionnaire | 118 |
| 6.7 | Pilot study..... | 118 |
| 6.8 | Ethical procedures | 120 |
| 6.9 | Data collection procedure | 120 |
| 6.9.1 | Measures of complexity..... | 122 |
| 6.9.2 | Measures of accuracy..... | 123 |
| 6.9.3 | Measures of fluency | 124 |
| 6.9.4 | Inter-rate reliability | 126 |
| 6.10 | Data Analysis..... | 127 |
| 6.11 | Conclusion..... | 127 |
| 7. | Chapter 7: RESULTS: STUDY TWO..... | 128 |
| 7.1 | Introduction | 128 |

| | | |
|-------|---|-----|
| 7.2 | Preliminary analysis | 129 |
| 7.3 | Multivariate analysis of variance | 130 |
| 7.4 | Paired-samples t-test | 131 |
| 7.5 | Effects of IR on L2 oral performance | 133 |
| 7.5.1 | Effects of IR on syntactic complexity..... | 133 |
| 7.5.2 | Effects of IR on lexical complexity | 133 |
| 7.5.3 | Effects of IR on accuracy..... | 134 |
| 7.5.4 | Effects of IR on fluency..... | 134 |
| 7.6 | Effects of IR on perceptions of Task Difficulty..... | 135 |
| 7.7 | Effects of TC as mediated by language proficiency | 138 |
| 7.7.1 | LP and TC effects on syntactic complexity | 139 |
| 7.7.2 | LP and TC effects on lexical complexity..... | 141 |
| 7.7.3 | LP and TC effects on accuracy | 142 |
| 7.7.4 | LP and TC effects on fluency | 144 |
| 7.8 | Effects of TC as mediated by working memory | 147 |
| 7.8.1 | WM and TC effects on syntactic complexity | 148 |
| 7.8.2 | WM and TC effects on lexical complexity..... | 149 |
| 7.8.3 | WM and TC effects on accuracy | 151 |
| 7.8.4 | WM and TC effects on fluency..... | 152 |
| 7.9 | LP and WM as predictors of L2 oral performance | 155 |
| 7.10 | Summary of the key findings | 158 |

| | | |
|-------|--|-----|
| 8 | Chapter 8: DISCUSSION: STUDY TWO..... | 159 |
| 8.1 | Introduction..... | 159 |
| 8.2 | IR and L2 oral performance..... | 159 |
| 8.3 | IR and CALF measurements..... | 163 |
| 8.4 | IR and perceptions of task difficulty..... | 168 |
| 8.5 | IR and models of task complexity..... | 169 |
| 8.6 | IR and models of speech production..... | 171 |
| 8.7 | LP and TC interaction effects on L2 performance..... | 173 |
| 8.7.1 | LP-TC interaction effects on syntactic complexity..... | 174 |
| 8.7.2 | LP-TC interaction effects on lexical complexity..... | 176 |
| 8.7.3 | LP-TC interaction effects on accuracy..... | 177 |
| 8.7.4 | LP-TC interaction effects on fluency..... | 178 |
| 8.8 | WM and TC interaction effects on L2 performance..... | 180 |
| 8.8.1 | WM-TC interaction effects on syntactic complexity..... | 181 |
| 8.8.2 | WM-TC interaction effects on lexical complexity..... | 182 |
| 8.8.3 | WM-TC interaction effects on accuracy..... | 183 |
| 8.8.4 | WM-TC interaction effects on fluency..... | 184 |
| 8.9 | LP and WM predictability of L2 oral performance..... | 186 |
| 8.10 | Conclusion..... | 188 |
| 9 | Chapter 9: CONCLUSION..... | 189 |
| 9.1 | Introduction..... | 189 |

| | | |
|-------|---|-----|
| 9.2 | Conclusions from the findings | 189 |
| 9.2.1 | Conclusions from Study One | 189 |
| 9.2.2 | Conclusions from Study Two | 190 |
| 9.3 | Contributions of this study | 191 |
| 9.4 | Implications of this study | 195 |
| 9.5 | Limitations and suggestions for future research | 197 |
| 9.6 | Final remarks..... | 199 |
| | List of references | 200 |

List of Tables

| | |
|---|-----|
| Table 1. Robinson's (2010) SSARC Model for increasing TC..... | 33 |
| Table 2. The study design and variables of Study One | 68 |
| Table 3. Participants' scores in internal English exams and continuous assessments | 69 |
| Table 4. Measures of complexity, accuracy, and fluency in Study One..... | 79 |
| Table 5. Descriptive statistics for the dependent variables (Study One)..... | 83 |
| Table 6. Paired-samples t-tests and effect sizes (Study One) | 86 |
| Table 7. Lexical analysis of +IR vs –IR tasks (Study One)..... | 88 |
| Table 8. Thematic analysis of perceptions of TD (Study One) | 91 |
| Table 9. The study design and variables of Study Two..... | 112 |
| Table 10. Participants' scores in Oxford Placement Test..... | 114 |
| Table 11. Participants' scores in the elicited imitation task | 115 |
| Table 12. Participants' combined language proficiency scores | 116 |
| Table 13. The participants' overall language proficiency levels | 116 |
| Table 14. Participants' scores in the working memory tests | 117 |
| Table 15. The participants' working memory levels | 117 |
| Table 16. Measures of CALF in Study Two..... | 125 |
| Table 17. Descriptive statistics for the dependent variables of Study Two..... | 129 |
| Table 18. Paired-samples t-tests and effect sizes (Study Two) | 132 |
| Table 19. Participants' justifications of TD perceptions (Study Two)..... | 137 |
| Table 20. Descriptive statistics for syntactic complexity (LP as between-subjects variable) | 140 |

| | |
|---|-----|
| Table 21. Effects of LP and TC on syntactic complexity (Two-way ANOVA) | 140 |
| Table 22. Descriptive statistics for lexical complexity (LP as between-subjects variable).... | 141 |
| Table 23. Effects of LP and TC on lexical complexity (Two-way ANOVA) | 142 |
| Table 24. Descriptive statistics for accuracy (LP as between-subjects variable) | 143 |
| Table 25. Effects of LP and TC on accuracy (Two-way ANOVA) | 144 |
| Table 26. Descriptive statistics for speed fluency (LP as between-subjects variable) | 144 |
| Table 27. Effects of LP and TC on speed fluency (Two-way ANOVA)..... | 145 |
| Table 28. Descriptive statistics for filled pauses & repairs | 146 |
| Table 29. Effects of LP and TC on filled pauses and repairs (Two-way ANOVA)..... | 147 |
| Table 30. Descriptive statistics for syntactic complexity (WM as between-subjects) | 148 |
| Table 31. Effects of WM and TC on syntactic complexity (Two-way ANOVA)..... | 149 |
| Table 32. Descriptive statistics for lexical complexity (WM as between-subjects variable) . | 150 |
| Table 33. Effects of WM and TC on lexical complexity (Two-way ANOVA) | 150 |
| Table 34. Descriptive statistics for accuracy (WM as between-subjects variable) | 151 |
| Table 35. Effects of WM and TC on accuracy (Two-way ANOVA)..... | 152 |
| Table 36. Descriptive statistics for speed fluency (WM as between-subjects variable)..... | 153 |
| Table 37. Effects of WM and TC on speed fluency (Two-way ANOVA)..... | 154 |
| Table 38. Descriptive statistics for filled pauses and repairs (WM as between-subjects)..... | 154 |
| Table 39. Effects of WM and TC on filled pauses and repairs (Two-way ANOVA) | 155 |
| Table 40. Multiple regressions for LP and WM predicting oral performance | 156 |
| Table 41. Lexical analysis of +IR vs -IR tasks (Study Two)..... | 164 |

List of Figures

| | |
|--|-----|
| Figure 1. Monolingual speech production model (Levelt, 1989, p. 9) | 18 |
| Figure 2. The model of bilingual speech production (Kormos, 2006, p.168) | 20 |
| Figure 3. Robinson's TCF for task classification (Robinson, 2015, p.96)..... | 31 |
| Figure 4. The Multicomponent model of WM (Baddeley, 2003, p. 835)..... | 55 |
| Figure 5. Participants' perceptions of Task Difficulty (Study One)..... | 90 |
| Figure 6. Means of syntactic complexity performance (Study One) | 95 |
| Figure 7. Means of lexical complexity performance (Study One) | 96 |
| Figure 8. Means of accuracy performance (Study One)..... | 99 |
| Figure 9. Percentage of Task Difficulty rating (Study One)..... | 101 |
| Figure 10. Participants' perceptions of Task Difficulty (Study Two) | 136 |
| Figure 11. Percentage of Task Difficulty rating (Study Two) | 168 |
| Figure 12. Effects of LP and TC on syntactic complexity..... | 175 |
| Figure 13. Effects of LP and TC on lexical complexity | 176 |
| Figure 14. Effects of LP and TC on accuracy..... | 177 |
| Figure 15. Effects of LP and TC on speed fluency..... | 178 |
| Figure 16. Effects of LP and TC on filled pauses and repair..... | 180 |
| Figure 17. Effects of WM and TC on syntactic complexity..... | 181 |
| Figure 18. Effects of WM and TC on lexical complexity | 183 |
| Figure 19. Effects of WM and TC on accuracy | 184 |
| Figure 20. Effects of WM and TC on speed fluency | 185 |
| Figure 21. Effects of WM and TC on filled pauses and repairs | 186 |

List of Appendices

| | |
|--|-----|
| Appendix 1: Task instructions (English version) | 210 |
| Appendix 2: Task instructions (Arabic version)..... | 212 |
| Appendix 3: Ethics Forms | 214 |
| Appendix 4: Language Background Questionnaire..... | 216 |
| Appendix 5: Task Difficulty Questionnaire | 217 |
| Appendix 6: The coding symbols | 218 |
| Appendix 7: Examples of types of errors | 219 |
| Appendix 8: Examples of repairs | 219 |
| Appendix 9: Samples of the coded data (Study One)..... | 220 |
| Appendix 10: Oxford Placement Test | 226 |
| Appendix 11: Elicited Imitation Task..... | 236 |
| Appendix 12: Working memory tests (Arabic & English)..... | 237 |
| Appendix 13: Samples of the coded data (Study Two) | 239 |

List of Abbreviations

| | |
|---------------|--|
| CAF | Complexity, Accuracy and Fluency |
| CALF | Syntactic Complexity, Accuracy, Lexical Complexity and Fluency |
| CEFR | Common European Framework of Reference |
| CH | The Cognition Hypothesis |
| CLT | Communicative Language Teaching |
| EFC | Percentage of error free clause |
| EPHW | Errors per 100 words |
| EFL | English as a foreign language |
| H | Hypothesis |
| IL | Interlanguage |
| IR | Intentional Reasoning |
| L1 | First Language |
| L2 | Second Language |
| LAC | Limited Attentional Capacity |
| LP | Language Proficiency |
| LT | Language Teaching |
| MLASU | Mean length of AS unit |
| MLC | Mean length of clauses |
| MLECSP | Mean length of end-clause silent pauses |

| | |
|---------------|---|
| MLMCSP | Mean length of mid-clause pauses |
| NECSP | Number of end-clause silent pauses |
| NFP | Number of filled pauses |
| NMCSP | Number of mid-clause silent pauses |
| NR | Number of repairs |
| PSR | Pruned speech rate |
| ROS | Ratio of subordination |
| RQ | Research question |
| SLA | Second Language Acquisition |
| SPSS | Statistical Package for the Social Sciences |
| TBLT | Task-Based Language Teaching |
| TC | Task Complexity |
| TCF | Triadic Componential Framework |
| TD | Task Difficulty |
| TL | Target Language |
| UPSR | Unpruned speech rate |
| WCR | Weighted clause ratio |
| WM | Working Memory |

Abstract

A paramount discussion on the cognitive approaches to Task-Based Language Teaching (TBLT) is the issue of predicting the systematic effects of cognitive task complexity (TC) on second language (L2) performance and the factors that interact with the effects of TC (Baralt, 2013; Kormos & Trebits, 2012; Révész, Michel, & Gilabert, 2016; Robinson, 2007; Skehan & Foster, 1999; Tavakoli, 2014). Two competing models of TC, i.e. the Cognition Hypothesis (Robinson, 2001) and the Limited Attentional Capacity (Skehan, 1998) have informed this research agenda. However, there is still a need to more carefully define and systematically operationalise intentional reasoning (IR) as a TC variable within task-based research. More importantly, more research is needed to investigate the interaction between the effects of TC and L2 learners' individual differences.

This thesis draws on the findings of two inter-related studies. Study One aimed to investigate whether increasing TC through IR demands would be associated with an increase in syntactic complexity, lexical complexity, and accuracy, and a decrease in fluency of L2 learners' oral performance. This study further investigated whether the +IR task would be perceived as more difficult. IR was operationalised on two levels i.e., task instructions and task content. A mixed-methods within-participants study design was conducted with 20 Jordanian secondary school students who performed two video-based oral narrative tasks with varying degrees of IR and completed a retrospective questionnaire on their perceptions of task difficulty (TD). The design was counter-balanced to avoid any practice or order effects.

Following the analysis of the participants' oral performance which was operationalised through a number of CALF measures, the findings of Study One revealed a systematic positive impact of IR on syntactic complexity and accuracy, and a negative impact on lexical complexity. However, fluency was not significantly affected by the IR demands. The participants perceived the +IR task as more difficult than the -IR task. They further attributed the difficulty to the IR demands which were required by task instructions and to the unfamiliarity and unpredictability of the content of the +IR video clip. These mixed results acknowledged the need to consider a possible interaction between the learners' individual differences and the effects of IR demands on L2 speech production.

Study Two was then designed to examine: 1) the effects of manipulating TC by IR in oral narratives on learners' L2 speech performance and perceptions of TD; 2) whether learners'

individual differences in language proficiency (LP) and working memory (WM) mediate the effects of IR; and 3) to what extent LP and WM can predict performance on tasks of different degrees of TC. Employing a mixed-methods approach, the study had a 2 x 2 within-between-participants factorial design. The participants were 48 learners of English at a secondary school in Jordan. They performed the same two video-based oral narratives of Study One and completed a retrospective questionnaire to rate their perceptions of TD. A counter-balanced design was used to control for any impact of order or practice. Oxford Placement Test (Alan, 2004) and a set of elicited imitation tasks (Wu & Ortega, 2013) were used to measure the participants' LP, and backward-digit span tasks in L1 and L2 (Kormos & Trebits, 2011; Wright, 2010) were used to test their WM.

The participants' oral performance was analysed in terms of a number of CALF measures. The quantitative and qualitative data obtained from the questionnaire were carefully analysed. The results confirmed that IR demands resulted in significantly producing more syntactic complexity, accuracy, speed fluency, and filled pausing, whereas lexical complexity decreased in the +IR task. However, no effects were evident on silent pausing or repair fluency. The participants perceived the +IR task as more difficult than the -IR task. The same themes which were mentioned in Study One emerged as the main factors that contributed to the perceptions of TD, i.e. task-induced and task-inherent cognitive demands as triggered by task instructions and content.

Even though main effects were detected for LP and WM on some aspects of L2 performance, no interaction effects were significantly observed between TC and LP or WM. The findings further designated LP as a reliable predictor of speech performance with respect to lexical complexity, accuracy, speed fluency, and pausing fluency. However, WM did not statistically explain variations in any aspect of L2 oral performance but correlated significantly with accuracy and lexical complexity. These results imply that considering cognitive task complexity in isolation may provide a too simplistic picture of what is happening during task performance. Furthermore, the study highlights the importance of considering the interrelation between the cognitive demands of a task and its linguistic requirements to explain intentionality when making decisions on what analytic measures of CALF to employ. The findings have also substantial implications for L2 pedagogy and research.

Acknowledgements

This thesis could not have been completed without many supportive people around me. First, I would like to express my sincere gratitude to Dr. Parvaneh Tavakoli, my first supervisor, for her tremendous support, constructive comments and invaluable advice. Her super supervision, which has gone beyond any expectations, has made writing this thesis easier but made writing this section so difficult as I could not find the words that really describe her. I would like also to extend my deepest gratefulness to Dr. Clare Wright, my wonderful second supervisor, for her valuable contributions to this thesis and for her guidance and encouragement. I am greatly indebted to both of them for the time and efforts they have generously invested in guiding and enriching my work. I am also grateful to Dr. Jacqueline Laws, the Director of Postgraduate Research Studies, who has taught me the principles of research with patience and a touch of creativity. My appreciation is also extended to Professor Rodney Jones, the Head of the Department of English Language & Applied Linguistics (DELAL), for his encouragement and limitless support. I would like also to thank the staff of DELAL and University of Reading for being so supportive and for making my study experience unforgettable.

I am sincerely thankful for the students, teachers and management of Ridwan Schools in Jordan for their precious time and contribution in collecting the data. Finally, and most importantly, my sincerest gratefulness goes to my lovely wife, Liza and my three children (Malek, Enas, Masa) who have sacrificed so much to allow me to fulfil my dreams. They have inspired me to think big, dream big, and realise that impossible is nothing! This thesis is devoted to the memory of my parents who passed away long time ago. Thank you, my Lord, for all the blessings you have granted me and for all those supportive people around me.

Chapter one: INTRODUCTION

1.1 Rationale for the study

As a teacher of English as a foreign language (EFL), I noticed that when my students found it difficult to perform certain oral tasks (e.g., narrate a story, or engage in a debate), they often blamed their futile lexical repertoire for not being able to achieve the desired task outcome. While the importance of having a rich lexical repertoire or control over the morphosyntactic features of the second language for successfully performing a language task cannot be denied, there are task-related factors that can make L2 performance demanding and complex. Understanding the way task characteristics contribute to the difficulty L2 learners experience while performing a task and the impact of performing more cognitively complex tasks on learners' L2 oral performance are the main focus of this thesis, not to mention exploring the contributions of what learners bring to language tasks through their individual differences.

EFL Students and teachers in Arab countries like Jordan and Saudi Arabia where I have learned and taught English are also facing many constraints that **result** in humble learners' language performance and attainment of L2 communicative skills. Both teachers and learners find themselves work on language tasks and materials that rarely mirror the learners' needs, values, beliefs and cultures (Shehdeh, 2010). This leaves EFL teachers struggling to facilitate language tasks that do not satisfy their learners' needs. Moreover, many learners feel demotivated to get engaged in those pedagogic tasks because they are not attached to real life situations, not to mention lack of authentic use of the target language outside classroom (Rahman & Alhaisoni, 2013). This leaves teachers alone again to deal with demotivated, frustrated and resistant students (ibid). Other challenges this context is facing include teacher-centeredness, inflexibility of teaching materials, assessment and testing tools that lack validity and reliability, large-size classrooms, negligence of learners' individual differences, and the resistance to adopt new teaching approaches to replace the traditional ones (Al-Jamal & Al-Jamal, 2013; Rahman & Alhaisoni, 2013; Shah, Hussain, & Nasseef, 2013; Shehdeh, 2010).

The future of Task-Based Language Teaching (TBLT) as an approach to Second Language Acquisition (SLA) and language teaching (LT) could still be questioned in countries like Jordan and Saudi Arabia. It is where rote learning has roots for many generations when people used to memorise Quran and Prophet Mohammed's sayings to pass them to the next generations, not to mention the Arabic poetry and literature (Long, 2015b) . It could be still argued that rote

learning is still dominant and considered a key element in L2 classrooms in the Middle East countries until today (Elyas & Picard, 2010; Smith & Abouammoh, 2013). Therefore, Long (2015a) anticipates that “the relative spontaneity and creative, communicative language use in typical TBLT lessons could initially sit uneasily with the heavy reliance on rote memorization in education in countries as Saudi Arabia” (p.370). Added to this is the lack of systematic research investigating what other factors contribute to a deficient language performance and acquisition of a second language among learners in such countries (Shah, Hussain, & Nasseef, 2013).

This reality about learning and teaching EFL in that part of the world has inspired this research to attempt to broaden knowledge and understanding about the progressive end product of TBLT as a promising and effective approach to LT that can replace the traditional methods, and hence overcome any potential obstacles which threaten its future in these countries. Though, it is not a main aim of this study, it is still hoped that doing research within TBLT context, will bring more attention to this approach, and help develop a more in-depth understanding of TBLT as an effective approach to L2 learning and teaching in this part of the world.

The current study is motivated by the burgeoning interest in using task as a research and learning tool within TBLT. The tendency to adopt more cognitive approaches to SLA has made this study of a significant importance. Researching cognitive task complexity is of immense importance because it provides teachers and syllabus designers with information about the level of challenges a task should have to appropriately match learners’ proficiency level (Skehan, 1998). Skehan advocates that Task Complexity (TC) research could benefit L2 performance and development in two ways. First, tasks with appropriate level of complexity are more motivating for learners, and hence learners are likely to get more engaged in performing language tasks. Second, learners will be able to cope with task demands when challenged with attained levels of task requirements, and thus utilise and channel their attentional, cognitive and memory resources more effectively. As a result, more balanced performance on all language aspects and better opportunities for interlanguage development are more likely to take place.

Researching TC can help SLA and LT researchers understand how learners develop their interlanguage during performance, and hence acquire the target language (TL) more sufficiently. Given the significant role learners’ attentional resources play in language processing, performance and acquisition (Robinson, 2011a), researching task complexity would allow researchers to learn about which aspects of TC interacts with attentional resources and to what

extent TC can be manipulated to encourage an effective use of attentional resources. Robinson (2007) argues that the accumulated findings of TC studies can serve as empirical rationale for selecting, grading, and sequencing tasks within syllabus design and classroom implementation.

Tavakoli (2009b) suggests that identifying the factors that influence TC and perceptions of TD is essential for language teaching, syllabus design, and language testing. Getting an insider viewpoint about how task characteristics contributing to TC are perceived by learners “will broaden the current understandings of TD and will assist language educators in designing and employing more effective language teaching materials” (ibid, p.2). Therefore, there is a pressing need to clearly distinguish between TC and TD to solve the misunderstanding that has led to use these two constructs interchangeably by some researchers. Selecting an appropriate level of TC provides a more reliable and valid assessment of language performance, and consequently ensures more accurate testing results. Moreover, an in-depth exploration of the construct of TD can offer empirical evidence that helps to develop an index of TD which has been needed for a long time (Candlin, 1987; Nunan, 1989; Skehan 1998). Researching TD can also help researchers establish validity of TC manipulations and improve task design for more reliable and generalisable findings (Révész et al., 2016).

This study is further motivated by the recent calls to employ more systematic and analytic frameworks to investigate TC (Jackson & Suethanapornkul, 2013; Malicka & Sasayama, 2017). Moreover, the various ways task-based researchers conceptualise and operationalise TC variables in their studies have resulted in depicting inconsistent and less clear pictures regarding the impact of TC on both L2 performance and development. The abstractness of some of TC factors, the disagreement on their definitions, the practical issues that hinder data gathering, and the complications of the methodologies used in exploring TC factors are some of the challenges that face this line of research. Intentional reasoning (IR) which is employed in this study as a variable to operationalise TC is no exception.

The rationale behind investigating IR is justified by the lack of TC studies that explored the construct of IR. Moreover, previous studies that examined the effects of manipulating IR on L2 learners’ performance (e.g., Ishikawa, 2008; Robinson, 2007) were not successful in proposing convincing and coherent definitions of IR, not to mention the absence of systematic frameworks to operationalise IR, which can be effectively in the service of future research. With that in mind, this study is planning to address these limitations and bridging this gap in TC research. Therefore, this study is intended to re-conceptualise and re-operationalise IR construct more

systematically, and further attempts to make a novel contribution by offering a framework of IR operationalisation that can be employed in future studies to help harmonise research efforts in this kind of studies, and hence obtain more consistent findings with respect to IR and TC.

This study is also motivated by the paucity of research investigating the interaction between TC, TD and the individual differences between L2 learners. Research on the interaction between learners' individual difference and TC factors has either received less attention from researchers or has been considered a second aim in many studies (Robinson, 2015). Considering learner factors and what learners can add to task performance through their ability and affective variables is assumed to offer valuable data which can help in predicting the outcomes of language tasks, and hence informing their design, and ensure their effective implementation in L2 classrooms.

Language proficiency (LP) and working memory (WM) as individual variables, are assumed to play crucial roles in L2 processing and learning, and hereby affect both language performance and development (Skehan, 2015b). Previous research on TC has explored IR but has not considered its interaction with language proficiency and working memory. The question how individual differences, i.e. LP and WM interact with IR demands to affect L2 oral performance has not been answered yet. This study attempts to answer this question and contribute to advance research on SLA and LT. To the best of my knowledge, the relationship between LP and WM as learner factors and IR as a TC factor has not been explored before in TBLT context, which is a research gap this study is designed to address.

This study will be designed in a way that it does make original contributions to TBLT and SLA research through 1) the well-justified rationale mentioned earlier; 2) a mixed-method study design that incorporates quantitative and qualitative data; and 3) systematic conceptualisation and operationalisation of the constructs under investigation. The latter will be achieved through the attempt of this study to redefine and re-operationalise IR in a way that addresses the shortcomings of previous research. It is expected that the findings of this study will advance our knowledge and understanding on how TC can be manipulated to promote L2 learners' speech performance and development, taking into consideration its interaction with LP and WM. As a final fine product, this study is attempting to offer guidance for future TC research in a way that leads to more collaborative efforts that can advance research on TC in particular and TBLT and SLA in general.

1.2 Organisation of the study

This thesis is formed of two interrelated studies. Study One investigates the effects of manipulating TC along intentional reasoning (IR) demands on L2 learners' oral performance measured by syntactic complexity, accuracy, lexical complexity and fluency (CALF) and perceptions of Task Difficulty (TD) measured by a retrospective questionnaire. Study Two is further designed to address the recommendations of Study One, which suggests considering the role of L2 learners' individual variables in mediating and/or predicting speech performance on more complex tasks. Therefore, in addition to investigating the effect of increasing IR demands on CALF and perceptions of TD, Study Two is designed to examine whether the effects of increasing TC along IR requirements on L2 learners' oral performance are mediated by variations in L2 learners' language proficiency (LP) and working memory (WM). Furthermore, the study attempts to explore whether the language performance of tasks that require degree of IR demands can be predicted through variations of L2 learners' LP and WM.

This thesis is organised in nine chapters. This chapter has already presented the rationale for the study and the importance of this line of research. Chapter 2 reviews the relevant literature on task and TBLT, models of speech production, and aspects of language performance. The chapter then focusses on discussing the main variables under investigation and the theoretical background of this study. Definitions of TC are provided, followed by presenting the two theoretical models of TC, i.e. the Cognition Hypothesis and the Limited Attentional Capacity. Previous research on TC is then reviewed, before offering a detailed overview of IR as the main independent variable which will be employed to operationalise TC in this study. The Literature Review Chapter is then dedicated to the issue of individual differences in TBLT research. This leads to provide a full description of language proficiency and working memory as the individual variables that are included in Study Two. LP is defined, and then the measurements of LP are thoroughly discussed. The relationship between LP and language performance is explained, before reviewing previous studies that explored the effects of LP on L2 performance and development. The last section of this chapter is devoted to the construct of WM, which is defined and explained within Baddeley's (2000) model of WM. This is followed by explaining how WM is measured within SLA and TBLT context. Finally, the role of WM in L2 performance and development is described, before finally reviewing previous research that investigated the interaction between WM and language performance and development.

Chapter 3 is dedicated to presenting the methodology of Study One. The chapter begins by stating the aims of Study One, the research questions and related hypotheses. Second, detailed accounts of the study design, participants, tasks, materials, pilot study, and ethical procedure are provided. Third, data collection procedure is described in detail. Fourth, the twelve measures adopted to represent the four aspects of language performance, i.e. CALF are described and justified. The chapter ends with explaining the procedures of transcribing, coding and analysing the data.

Chapter 4 presents the results of Study One. First, the chapter outlines the descriptive analyses obtained for the twelve measures which operationalised the participants' speech performance. Second, the choice of running MANOVA analysis is justified, before presenting its results. Third, the use of paired-samples t-tests is also justified, and then a detailed description of the t-tests results is presented with respect to syntactic complexity, lexical complexity, accuracy, and fluency. Finally, the quantitative and qualitative results gained from the TD retrospective questionnaire are presented and explained.

Chapter 5 discusses the key findings of Study One. First, a summary of the main results is outlined. Second, the results of the effects of IR on each aspect of speech performance, i.e. CALF are thoroughly discussed in relation to the research questions, hypotheses and previous studies. Third, the findings regarding the participants' perceptions of TD are also discussed and matched to the relevant research question and prediction, before comparing the results with previous research on TD. The issues that emerged with regards to the choice of CALF measurements are exposed to a careful discussion. The final remarks of Study One are projected, before a proposal for Study Two is presented to defend the rationale for designing a new study and make a link between Study One and Study Two.

Chapter 6 presents a detailed description of the methodology employed in Study Two. First, the aims of Study Two are clarified, before presenting the research questions and relevant hypotheses. Second, the design of the study is explained and justified. Third, the participants, tasks, materials, pilot study, and ethical procedure are presented in detail. Fourth, the procedure of data collection is described. Five, the twenty measures selected to operationalise language performance are presented and their choices are justified. Finally, the procedures of transcribing, coding, and analysing data are presented.

Chapter 7 is devoted for presenting the results of Study Two. First, the chapter summarises the descriptive analyses attained for the twenty measures of CALF. Second, the results of running MANOVA are presented. Third, the findings achieved from running the paired-samples t-tests are described to answer the research questions regarding the effects of IR demands on speech performance. Fourth, the quantitative and qualitative results obtained from the TD retrospective questionnaire are explicated to answer the research question about the participant's perceptions of TD. The next section describes the results of the ANOVA analyses regarding the interaction effects of LP and TC on speech performance, followed by presenting the ANOVA results with regards to WM-TC interaction effects on speech performance. The last section is dedicated to presenting the results of the regression analyses to answer the final research question on whether LP and WM are reliable predictors of speech performance on tasks with increased TC.

Chapter 8 is dedicated to explaining and discussing the main findings of Study Two. The first section outlines the key findings presented in the Results Chapter. The second section discusses the results regarding the effects of IR on CALF in relation to Research Question 1, and then the findings are compared to those achieved in previous studies, before discussing the issues of CALF measurements that emerged from the findings. The third section is devoted to discussing the results achieved from the TD questionnaire with reference to the predictions of Research Question 2 and the findings of previous research about L2 learners' perceptions of TD. The fourth section is used to interpret and discuss the results in association with the models of TC and models of speech production. The fifth section is devoted to discussing the ANOVA results obtained to answer Research Question 3 which questions whether there is an interaction between LP and TC. The findings are interpreted with respect to the predictions and the findings of previous research. The sixth section is dedicated to discussing the ANOVA results achieved to answer Research Question 4 which enquires any possible interaction between WM and TC. The findings are again interpreted in association with the predictions and the findings of previous research. The last section discusses the results of the regression analyses which are run to answer Research Question 5 concerning LP and WM predictability of learners' speech performance on more complex tasks.

Chapter 9 draws conclusions from the key findings of Study One and Two, and emphasises their significance. The contributions of this study are then highlighted, before discussing their potential implications. The limitations of Study One and Study Two are then acknowledged. Recommendations and suggestions for potential areas for future research are finally made.

Chapter 2: LITERATURE REVIEW

2.1 Introduction

Studying task complexity (TC) has recently become a central focus of Task-Based Language Teaching (TBLT) research, and has attracted substantial attention among researchers around the world (Baralt, 2013; Gilabert, 2007a; Pallotti, 2015; Robinson, 2011a; Skehan, 2015a; Tavakoli, 2014). Within psycholinguistic and cognitive perspectives, task has been employed as a unit of investigation as it enables researchers examine processes involved in language performance. The findings of this body of research, mainly emerging from measuring language performance in terms of its syntactic complexity, lexical complexity, accuracy and fluency, aim to identify variables that affect performance, accelerate or hinder learning processes, and hence influence the effectiveness of L2 learning and teaching within TBLT. Furthermore, these efforts endeavour to bridge the gap between pedagogy and research, and between LT and SLA.

This chapter has three main sections, which includes first a discussion of task and TBLT. Second, introduction and discussion of theoretical frameworks that inform this study which includes models of speech production and models of TC. The next section discusses the variables of the study which includes presenting intentional reasoning (IR) as a factor that operationalises TC in this study. This chapter further gives an account of the individual differences and task performance before presenting language proficiency (LP) and working memory (WM) which will be included as independent variables to the investigation in Study Two.

2.2 Definitions of Task

Following the tendency to employ more communicative approaches to LT, a need has emerged to bring real life situations to the instructional environments in which L2 learners will be involved in learning language by modelling tasks that resemble those outside classrooms (Willis, 1996). These new demands and high expectations from L2 learners, teachers and other stakeholders have helped task to find its own way into language teaching and learning contexts as a vehicle to communicate in the target language (TL) more effectively. This meaningful use of TL is assumed to trigger associated linguistic representations and acquisitional processes, and henceforward enrich and/or stabilise the learners' interlanguage (IL) (Robinson, 2010). It is therefore no wonder that task has become a focus of a considerable amount of research in the

contexts of TBLT and SLA. Within the context of TBLT research, task has been employed mainly “as a vehicle to elicit language production, interaction, negotiation of meaning, processing of input, and focus on form” (Van den Branden, 2006, p. 1).

Lack of consensus among researchers on what a *task* is has made defining task difficult. This could be attributed to the tendency of some researchers to adapt and adopt definitions that serve their narrow research interests, which may cause inconsistency on how to design, manipulate and research task. Researchers distinguish between a target task which is a real-life task outside the classroom, and a pedagogic task which comprises using language communicatively in classroom. Nunan (1989) defines a pedagogic task as “a piece of classroom work which involves learners in comprehending, manipulating, producing, or interacting in the target language while their attention is principally focused on meaning rather than form” (p.10). This definition advocates the need for form-meaning association to achieve the desired outcome of task. According to Nunan, task is learner-centred in the sense that learners are free to rely on the available forms and manipulate the target language to overcome any communication breakdown. However, the tendency for learner centeredness in task performance may diverge task from achieving its objectives or help learners stretch their IL. Supporting this point, Breen (1987) considers task as an operational activity that involves language and “has a particular objective, appropriate content, a specified working procedure, and a range of outcomes” (p.23). This broader definition emphasises the role of task as a workplan which aims to facilitate the process of language learning but in a more restricted fashion in terms of task goals and procedures.

Candlin (1987) sees tasks as sequenced classroom activities that are constructed cooperatively by teachers and learners which need to pose challenges. This description accentuates the importance of sequencing tasks but does not elucidate explicitly whether this sequencing needs to be complexity-oriented or outcome-oriented. Though problem-solving tasks can be advantageous to motivate speakers to be more productive, it could be also argued that learners can successfully communicate without a need for a problem to solve. Swales (2009) argues in the same direction and advocates that task is “one of a set of differentiated, sequenceable goal-directed activities drawing upon a range of cognitive and communicative procedures” (p.48). Swales agrees with other researchers on the value of presenting sequenced tasks but considers being goal-oriented as more advantageous than being problem-posing for a more successful task implementation. Consistent with this definition, Willis (1996) defines a pedagogic task as

“a goal-oriented activity in which learners use language to achieve a real outcome” (p. 53). This definition presents task as a vehicle to generate a natural purpose to use TL and pinpoints again the importance of learner-centeredness in approaching tasks. In addition to being goal-oriented, Long and Crookes (1992) debate that task should focus on meaning, which reflects a real need to use language naturally. Despite agreeing on the essential features of task, the definitions discussed above failed to depict a holistic picture that could offer in-depth descriptions of the multifaceted nature of task.

Highlighting its main features, Skehan (1998) defines task as “an activity in which: meaning is primary; there is some communication problem to solve; there is some sort of relationship to comparable real-world activities; task completion has some priority; and the assessment of the task is in terms of outcome” (p.95). Based on this definition, pedagogic tasks are not identical real-life tasks but instead are adapted to become applicable to classroom contexts. This leaves doors wide open to the possibility to redesign and manipulate task pedagogically to increase its effectiveness, appropriateness, and essentialness. Skehan further brings stakeholders’ attention to the importance of using the outcome of task as a tool to assess any gains regarding the elicited speech performance, and hence use it as a tool to measure learners’ proficiency level in the TL.

Following Skehan, R. Ellis (2003) describes task as a workplan that requires learners to use language pragmatically to achieve non-linguistic outcomes, and that “tasks are activities that call for primarily meaning-focused language use” (p.3). In his definition, Ellis makes a clear distinction between tasks which are meaning-focused that involve pragmatic meaning, and exercises which are form-focused that reflect semantic meaning. Building on this definition, Ellis suggests that a pedagogical task involves: a) a focus on pragmatic meaning; b) some kind of gap (e.g., *reasoning gap*); c) linguistic resources to complete the task; d) cognitive processes; e) communicative results; and f) any of the language skills. (ibid, p.11).

Skehan (1998) and R. Ellis (2003) seem to define task more adequately by including the elements that reflect the multidimensional nature of task in terms of focus (meaning), design (challenging) and outcome (communicative). Additionally, the two definitions give prominence to the role attention and memory can play in language performance, as task is supposed to impose some cognitive demands and challenges on learners (e.g., reasoning, problem solving). These demands are assumed to motivate learners to employ and combine their cognitive and linguistic resources to achieve the required task outcome. However, these definitions can still be criticised for not considering task from learners’ perspective, ignoring the fact that learners

may have other agendas rather than focusing on meaning or form while performing tasks. Furthermore, it is more likely that learners will not totally ignore structure while performing meaning-oriented tasks, even if they are not well-equipped linguistically. Furthermore, Long (1998) argues that, surprisingly, learners still have some concerns about form when they are pushed to focus on meaning only. For that reason, it could be argued that without a balance between meaning and form, sustained development in L2 might not be effectively facilitated or achieved. For the purpose of this study, Skehan's (1998) definition will be adopted for task, i.e. "an activity in which: meaning is primary; there is some communication problem to solve; there is some sort of relationship to comparable real-world activities; task completion has some priority; and the assessment of the task is in terms of outcome" (p.95).

2.3 Task-Based Language Teaching (TBLT)

TBLT emerged three decades ago in response to the increasing needs for more functional approaches to L2 learning and teaching. TBLT was an attempt to identify and satisfy "diverse communicative L2 needs in a rational, sufficient, psycholinguistically defensible manner" (Long, 2015a, p. 2) . Supported theoretically and empirically, TBLT initially emerged as a new trend of the communicative language teaching approach (CLT), to assist learners in using the target language in real-world communicative tasks, and hence advance L2 proficiency (R. Ellis, 2003). Prabhu's (1987) Procedural Syllabus was one of the first attempts to implement TBLT by developing a school project based on tasks sequenced according to level of complexity (problem solving), with learners acquiring English by negotiating for meaning only. Despite receiving a lot of criticism, Prabhu's Procedural Syllabus constituted a point of departure for proposing improved versions of TBLT. Following the Procedural Syllabus, Long and Crookes (1993) proposed a task-based teaching programme that was built on target tasks designed to be authentic, based on learners' needs, and sequenced for their TC. Complexity was manipulated by increasing the number of steps, solutions or elements required to do a task, in addition to the amount and type of language needed to perform the task.

The abovementioned earliest TBLT attempts have been envisioned to "guide the design and delivery of key components of any language educational programme, from teacher practice to curriculum to materials to assessment" (Norris, 2015, p. 30), and hence these attempts have influenced the development of TBLT globally. Since 2005, an international conference has been devoted solely to topics related to TBLT. The International Conference on TBLT aims to

establish collaborative efforts of language teachers, syllabus designers, and researchers around the world to highlight the major theoretical principles and practices of TBLT and to encourage the field to address the challenges that face language teaching (Bygate, 2015). Thus, many researchers still dedicate most of their efforts to investigate issues related to TBLT (Bygate, 2015; Gilabert & Barón, 2013; Norris, 2015; Nunan, 2006; Robinson, 2011b; Skehan, 2014; Tavakoli, 2009a).

2.3.1 Definitions of TBLT

The official website of the International Association for Task-Based Language Teaching (IATBLT) defines TBLT approach as “an educational framework for the theory and practice of teaching second or foreign languages. Based on empirical research, TBLT adopts meaning-based, communicative tasks as the central unit for defining language learning needs, determining curriculum goals, designing activity in language classroom, and assessing language competencies” (IATBLT, 2015). The definition considers the curriculum goals and syllabus activities as an outcome of learners’ needs analysis which might guarantee identifying more appropriate tasks that suit learners, and hence boost the validity and reliability of this approach. Moreover, this definition ascertains that the methodology of this approach stems from the related theories on how learners acquire or learn an L2 sufficiently. A strong version of TBLT is assumed to “offer a rationale and process for the implementation of language educational programs, including needs analysis, syllabus and materials design, instructional practice, learning assessment, and teacher development” (Norris, 2015, p. 27). Employing tasks that are relevant to L2 learners’ developmental needs, and are built on methodological and pedagogical selections that are attuned to SLA theories and practices, can contribute to brand TBLT as a leading approach that facilitates sufficient language use, and thus results in more successful language learning.

Van den Branden (2006) defines TBLT as “an approach to language education in which students are given functional tasks that invite them to use language for real-world, non-linguistic purposes” (p.1). Henceforward, TBLT provides improved settings for triggering learners’ acquisition processes, resulting in positive gains with respect to their language performance (Shehadeh, 2005). Given that, task is a key element in classroom because it constitutes the tool that pushes L2 learners to deploy their available linguistic and cognitive resources to accomplish the communicative outcomes of task. TBLT advocates for learners’ dynamic role in taking risk and experimenting new language as a prerequisite to L2 development

(Willis, 1996). TBLT can be also seen as an approach of practical language learning, which aims to assist learners to understand, process, internalise, and use new information within its meaningful contexts (Bygate, 2015). In this sense, learners are required to use the target language with a focus on meaning and develop non-linguistic understanding of the processed information resulting in directing their attention to attend form to express meaning. Accordingly, TBLT is an analytic approach to L2 learning and teaching with a focus on form during meaning-oriented tasks (Long, 2015b). Building on this, TBLT is rather an improved version of analytic approaches but with focus on form, which attempts to address the shortcomings of approaches which focus purely on meaning or on forms.

2.3.2 Emergence of TBLT

As noted above, the shortcomings of previous approaches and the increased demands for stronger versions of functional approaches to L2 learning and teaching have laid the foundations for TBLT as an innovative approach to LT and SLA. In the last few decades, there was a move towards more holistic, student-centred, and communication-based models of language learning and teaching to replace the traditional versions, motivated by the growing needs for more purposeful use of language (Van den Branden, Bygate, & Norris, 2009). This move was also supported theoretically by proposing for a vital role of language procedural knowledge in promoting acquisition and learning through meaningful practice of the TL (Skehan, 1998).

In reaction to the growing needs of L2 learners, inadequate traditional approaches and instructional practices, and the evolutionary theories of SLA and neighbouring disciplines, CLT approach emerged to keep LT on the same pace of development (J. Richards & Rodgers, 2014). CLT positioned the functional use of target language and classroom meaningful interaction as central to LT (Brumfit, 1984). Considering learners as negotiators and teachers as facilitators, CLT stressed on using the available communicative strategies to negotiate for meaning to promote fluency through pair and group work (J. Richards & Rodgers, 2014). Owing to different interpretations and implications of CLT around the world, many versions were employed to satisfy variation in educational policies, syllabus designs and L2 learning needs. A continuum of models emerged between ‘weak CLT’, which aimed to offer L2 learners with opportunities to use language functionally in order to achieve communicative competence, and ‘strong CLT’, which aimed to trigger the developmental processes of language learning in order to attain a control on form (Howatt, 1984). Cohered with the basic principles of CLT, TBLT found its way as a new and well-established approach to LT that adopted a strong version to

CLT. The importance of functional use of TL through performing meaning-oriented tasks that are linked to real life, is the foremost principle TBLT and CLT share. However, they differ in terms of syllabus design, the focus each model, and how tasks are designed and employed.

TBLT was then introduced as “a model of L2 education that was systematically conceptualized along holistic, meaning-focused, learner-driven lines” (Van den Branden et al., 2009, p. 5). TBLT, as its name suggested, employed task as a basic unit at the stages of syllabus design, methodology, and assessment. Van den Branden (2006) identified three types of tasks to respond to each stage: a) *target tasks* to respond to learners’ needs analysis; b) *pedagogical tasks* to promote learners’ language proficiency; and c) *assessment tasks* to test the learners’ language performance. Likewise, Long (2015a) postulated that recycling task at all levels, i.e. *syllabus design, implementation, and assessment* helped TBLT to, 1) meet learners’ needs; 2) achieve reliable evaluation of learners’ learning; 3) accomplish consistency with SLA research findings and theoretical perspectives; and 4) guarantee a promising future as a leading approach to LT.

2.3.3 Future of TBLT

Similar to all communicative approaches to LT, TBLT aims to “promote learners’ ability to use the target language in real communication” (Van den Branden, 2006, p. 2). It could be argued that TBLT is central to SLA and LT research but not yet instruction. Still, some researchers are more optimistic in their speculations about the future of TBLT. Willis (1996) advocated for a possible smooth and worthwhile transition from current traditional approaches to TBLT by making learners “understand both the principles behind it and the purpose of each component. Learners need to understand why it is different, and how they are likely to benefit in the long run” (p.138). Willis called also for encountering the learners’ resistance for change through isolating the problems that cause such resistance and suggesting practical solutions for each issue. Perceiving TBLT as an innovative approach could further face confrontation from teachers as it may challenge or threaten their existing teaching theories and practices. In order to solve this problem, R. Ellis (2003) suggests a partial implementation of TBLT through employing a weak version, i.e. *a task-supported model*, which could be acknowledged “unthreatening as it requires only a modification to the way teachers teach, rather than a radical change” (ibid, p.323).

Nonetheless, the future of TBLT as an innovative approach to LT is still questionable and sceptical (Long, 2015b). The challenge that faces task-based instruction is how to replace the traditional approaches which are still dominating in some parts of the world like task-related or grammar-based approaches. According to Shehadeh (2005) “the persistence of grammar-based instruction in many teaching contexts in the world, despite its relative failure to produce effective language users, is partly due to the fact that it creates conditions where teachers feel secure as they can predict the language that will be needed and they feel comfortable in their roles as knowers” (p.28). In order for an approach to LT to be widely applicable and sustainable, it needs to include ready-made teaching materials (J. Richards, 1984) and a framework which fits into classroom setting, teacher training, and course evaluation (Skehan, 1998). More robust empirical evidence confirming that TBLT results in better proficiency gains than the current methods is a prerequisite to propose that TBLT may be a reliable alternative approach to successful LT.

Despite being branded in many parts of the world as an innovative approach to LT and SLA, Long (2015b) doubts any near future domination of TBLT worldwide due to missing elements or negative factors that hamper change or smooth transition from current approaches to pure task-based ones. Long argues that TBLT requires greater teachers’ command of the TLs, more teacher training, and more technology resources. As a result, the availability of financial resources and organisational support are fundamental for any successful implementation of TBLT, which are not easy to offer in many parts of the world. A probable clash with some other educational cultures where rote learning is still the norm or with other educational systems which persist dictate and centralised perspectives could be additional obstacles that threaten trading TBLT as a world-wide approach to SLA and LT (ibid). This can be true in the Middle East where I have learned and taught EFL for many years. Stakeholder still have more reasons to resist rather than to accept replacing TBLT with the current approaches.

2.3.4 A Cognitive approach to TBLT

TBLT has been widely investigated from a cognitive processing approach to language learning and teaching (Skehan, 1998, 2014), an approach that offers a theoretical rationale for this study. Building on propositions of cognitive psychology, the cognitive approach to language learning suggests that cognitive processes (e.g., speech processing, information processing, encoding and decoding) which take place during language learning play a leading role in promoting acquisition and learning. Skehan (1998:95) claims that “transacting tasks will engage

naturalistic acquisitional mechanisms, cause the underlying language system to be stretched, and drive development forward.” Within a cognitive perspective, language learner is perceived as an 'information-processor' who attempts to utilize accessible cognitive resources (e.g., attention and memory) to successfully process language to achieve a specific communicative outcome (Skehan, 2015a). Such attention-based and memory-based standpoints to TBLT research have created a tendency to theorise about the capacity of these cognitive resources and explore whether they interact with language performance under specific task characteristics or performance conditions. Hence, it is inevitable for TBLT research to adopt a psycholinguistic approach as a “starting point for a credible analysis of the psycholinguistic processes involved in second language speaking” (Skehan, 2014, p.4).

Skehan (1998) posits that processing language involves three stages “input, central processing, and output, as well as the interaction between these stages,” (p.43). While noticing and regulating attention are crucial during the *input* stage, the second stage, i.e. *central processing*, requires learners to apply either an ‘*exemplar-based system*’ or a ‘*rule-based system*’ (ibid). The former option is a lexical and memory-based system that relies on ‘*pre-fabricated sequences*’ of language. Adopting the *exemplar-based system* is assumed to facilitate accessing and processing chunks more quickly and effectively, resulting in positive gains in term of fluency. On the other hand, the *rule-based system* entails restructuring and generating the underlying rules of TL, which is more demanding in terms of processing, and hence affects the accuracy and complexity positively at the expense of fluency. Moving between the two systems, learners are assumed to acquire the TL through restructuring the linguistic patterns within the rule-based system, and accumulating formulaic chunks within the exemplar-based system.

Investigating TBLT from a cognitive viewpoint (e.g., Albert, 2011; R. Ellis, 2009a; Foster & Skehan, 2013; Kormos & Trebits, 2012; Tavakoli, 2014), researchers have been motivated to find out how manipulating task characteristics, conditions and designs: 1) promote L2 performance in terms of complexity, accuracy, and fluency; 2) assist learners to develop their IL systematically and communicate effectively; 3) exploit focus on form through manipulating attention; and 4) interact with learners’ individual differences and affective factors to affect task performance. Furthermore, cognitive-oriented research on TBLT attempts to explain how tasks could be sequenced and manipulated based on increased cognitive complexity to achieve a more balanced L2 performance and development.

However, this line of research is not without any criticism or controversy. Cognitively-driven studies can be criticised for their inconsistency in: 1) defining task; 2) operationalising cognitive task complexity; 3) using measures that are sensitive to task manipulation; and 4) interpreting the findings within the models of speech production. Additionally, a cognitively-oriented research paradigm still undervalues learners' individual factors as key factors that influence language performance and development. Since this study is framed within a cognitive approach to TBLT, it will strive to consider the aforementioned criticisms and contribute to resolve such discrepancy by offering a more systematic definitions and operationalisations of the variables under investigation, careful selection of the tasks and measurements of analysis, and in-depth interpretations and discussions of its findings in association with two prevailing models of task complexity, i.e. Robinson's Cognition Hypothesis (2001) and Skehan's Limited Attention Capacity (1998) and two models of speech production, i.e. Levelt (1989) and Kormos (2006).

2.4 Models of speech production

Within a psycholinguistic perspective to SLA research, models of speech production, i.e., Kormos, 2006; Levelt, 1989, play a dynamic role in understanding the mechanism and processes of learners' L2 performance. While Levelt's model was proposed for L1, Kormos's model is an attempt to develop one for L2. However, both are still at model stage and subject to further investigation. The two models speculate about the complex nature of speech production by describing how speakers encode or formulate ideas into speech (Kormos, 2011), and therefore provide researchers with interpretations on how speakers plan, process and produce language under different task characteristics and conditions (Skehan, 2014). It is hence essential to thoroughly understand to what extent the phases of L1 and L2 speech from intention to production are similar or different.

2.4.1 Levelt's (1989) monolingual speech model

Levelt (1989) puts forward a psycholinguistic model to describe and explain the mysterious process that enables language speakers to "transform intentions, thoughts and feelings into fluently articulated speech" (p. 1). As illustrated in Figure 1 below, Levelt (1989) proposes that L1 speech production involves four interrelated stages, i.e. Conceptualisation, Formulation, Articulation and Self-monitoring.

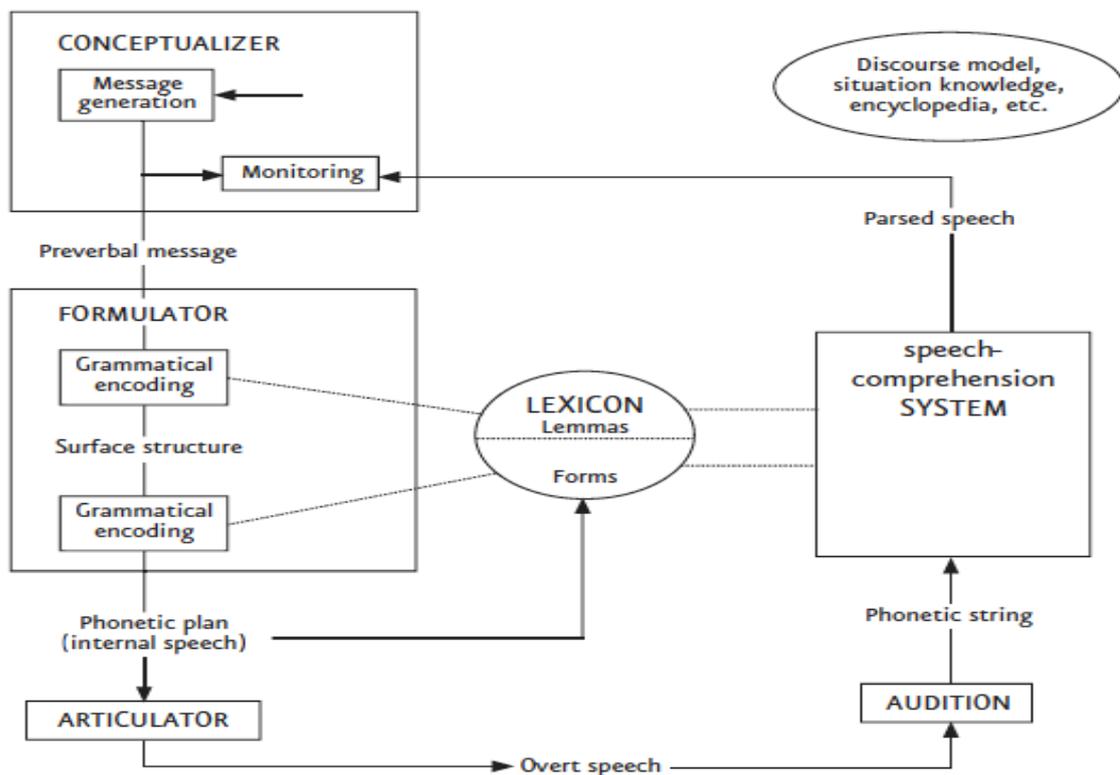


Figure 1. Monolingual speech production model (Levelt, 1989, p. 9)

2.4.1.1 Conceptualisation

Levelt (1989) describes the Conceptualisation as the preverbal message stage where the intended messages are formulated in the form of thoughts, ideas and propositions. This stage involves two sub-stages: First, it involves macro-planning which involves elaborating the goals of the intended message into sub-goals, and then planning the required speech acts, relying on the learner's procedural knowledge. The second sub-stage involves micro-planning, which shapes each planned unit with relevant information that is stored about the unit. The output of this process, i.e. the '*preverbal message*' is the input for the next stage.

2.4.1.2 Formulation

Levelt (1989) defines Formulation as translating the conceptual, preverbal information into linguistic components by accessing the appropriate grammatical, lexical, and phonological structures to encode the preverbal message. This requires speakers to access the mental lexicon in their long-term memory to encode: 1) *lemmas* which include semantic and syntactic information; and 2) *lexemes* which include phonological and morphological information. The product of this stage, the '*phonetic plan*' becomes the input to the next stage.

2.4.1.3 Articulation

Articulation process involves the execution of the phonetic plan through the organs of the neuromuscular system and transfer it into overt language. First, the articulatory plan is stored temporarily in the '*articulatory buffer*' device. Then, "the Articulator retrieves successive chunks of internal speech from this buffer and unfolds them for execution" (Levelt, 1989, p. 13). After being retrieved by the articulator for the execution, the final product, known as '*overt speech*' is produced by the articulatory organs.

2.4.1.4 Self-monitoring

Levelt (1989) proposes that a speaker is his/her own listener, therefore, speakers can monitor their preverbal and verbal message, while monitoring others' speech simultaneously. This audition process helps speakers to detect errors before or after execution. As a result, episodes of self-correction or scaffolding of others overt speech are generated as final products of the Self-monitoring stage.

Drawing on Levelt's model of speech production, R. Ellis (2003) proposes that when the processing conditions are demanding, L2 speakers will find it difficult to sustain this process, and as a result, accuracy suffers. On the other hand, when the conditions are less demanding, the L2 speaker will be able to give attention simultaneously to the Conceptualiser and Formulator, and therefore achieving a greater level of accuracy. Skehan (2009), agrees with Ellis that certain task characteristics can affect oral performance by manipulating the cognitive demands at the Conceptualization stage, versus characteristics that affect the Formulator. Skehan identifies four variables, i.e. *complexification*, *pressuring*, *easing*, and *focusing* as influential on the various stages of Levelt's model. An example of a task with higher cognitive demands that can have complexifying/pressuring influences is performing a retelling monologic task under time pressure. On the other hand, performing a dialogic task which involves concrete information while offering planning opportunity is less demanding, and hence can have easing/focussing effects (ibid). Skehan claims that complexification is linked mainly to the Conceptualisation stage and then to structural and lexical complexity, whereas pressuring, easing, and focusing are more relevant to the Formulator, and then, to accuracy and fluency. These proposals draw attention to the differences between L1 and L2 speech processing which demonstrate the need for a bilingual speech model to solve the puzzle of how learners process their L2 speech differently from their L1 under certain circumstances.

2.4.2 Kormos's (2006) bilingual speech model

Kormos (2006) has proposed a model of bilingual speech production based on Levelt's monolingual model. As shown in Figure 2 below, Kormos extends Levelt's monolingual model into a bilingual model in order to consider the differences in L1 and L2 speech production. She argues that her model is similar to an L1 model because it follows the same processing order but with some variations. Kormos argues that in L1, message planning requires attention, whereas formulation and articulation are automatic. Parallel speech processes are possible for native speakers because the required syntactic and phonological encoding draws on automatized linguistic knowledge, which can be accessed and executed very quickly, resulting in smooth and fast speech, unlike L2 speech.

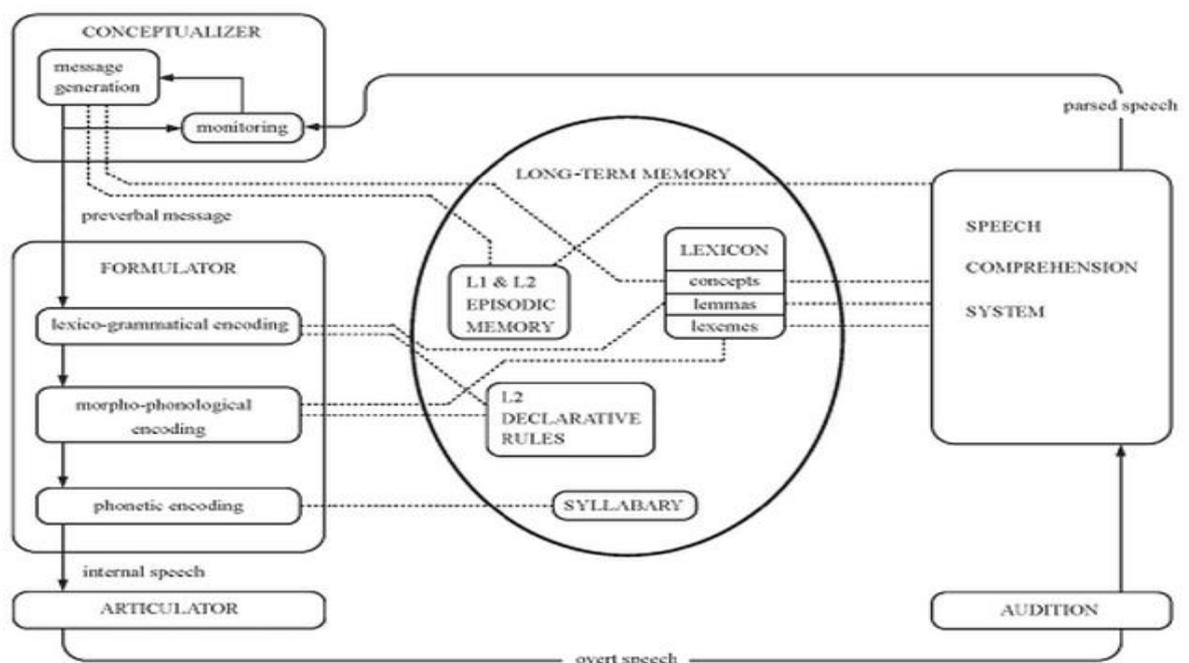


Figure 2. The model of bilingual speech production (Kormos, 2006, p.168)

Kormos (2006) explains that the two models share three knowledge stores: a) knowledge of the world; b) the mental lexicon; and c) the gestural scores. Kormos proposes a fourth store for L2 specific knowledge which contains a declarative memory of syntactic and phonological rules in the L2. Thus, the key features of L2 speech process that differ from L1 are summarized as:

- 1) The use of a declarative knowledge (not procedural).
- 2) The use of controlled processing (not automatic).
- 3) The use of serial processing (not parallel).

Kormos (2011) brings more attention to “the need for modified and extended models for L2 speech production in order to account for the differences between L1 and L2 production, namely the differences between L1 and L2 knowledge and the different nature of some of the L2 processing mechanisms such as syntactic and phonological encoding” (p. 42). Thereby, significant efforts have been devoted to developing the existing models of L2 speech production, by researching how L2 speakers process their speech cognitively and linguistically under different conditions. Measuring learners’ language performance in term of its syntactic complexity, lexical complexity, accuracy and fluency offers valuable data that enable this line of research to understand the multifaceted nature of L2 processing.

2.5 Aspects of speech performance

In the last two decades, complexity, accuracy and fluency (CAF) have emerged as a triad framework to investigate and measure L2 task performance and proficiency (Skehan, 2014). CAF, known recently as CALF by separating the construct of complexity into syntactic and lexical complexity have been widely employed as dependent variables to measure variation with respect to independent variables such as task features or performance conditions (Pallotti, 2009). Skehan (1996) is one of the first pioneers to coin CAF together in task-based research, suggesting that these dimensions can be measured separately and can reflect L2 development and proficiency. CALF are still considered “multilayered, multifaceted and multidimensional constructs” (Housen, Kuiken, & Vedder, 2012, p. 5), and hence there is a need to explore how these dimensions interrelate and interact during task performance, otherwise, research findings might miss the holistic picture of interlanguage development (Larsen-Freeman, 2009).

Housen and Kuiken (2009) indicate that CALF as a model is often used in SLA and TBLT studies to describe, quantify and gauge L2 learners’, 1) oral and written language production; 2) language proficiency; and 3) progress in acquiring or learning TL. Following the dominance of more cognitive approaches to task-based research, CALF have paved their way to the heart of TBLT investigations as dependent research variables to reflect “the psycholinguistic processes and mechanisms underlying the acquisition, representation and processing of L2 systems” (Housen et al., 2012, p. 2). Empirical evidence backs CALF as distinctive factors of L2 speech production (Norris & Ortega, 2009; Tavakoli & Skehan, 2005), and theoretical rationale justifies these aspects as sensitive factors that reflect modifications in the L2 system (Skehan, 1998). Accordingly, defining CALF thoroughly and understanding what each

dimension measures are essential before any claims or conclusions are to be advocated regarding learners' speech production. However, the different definitions L2 researchers have coined for each aspect, result in discrepancy in operationalising CALF, and hence affect the reliability of their findings.

As noted before, agreeing on one definition for each construct of CALF is not an easy task. Complexity is argued to be “the most complex, ambiguous, and least understood dimension of the CAF triad” (Housen & Kuiken, 2009, p. 463). Complexity can be interpreted in terms of both cognitive complexity which is L2 learner-user oriented, and linguistic complexity which is L2 system oriented (ibid). Despite postulating linguistic complexity as a main component of cognitive complexity, both facets of complexity are rated and operationalised differently. Cognitive complexity is supposed to be more subjective as it is linked to the learners' perceptions of difficulty. On the other hand, linguistic complexity is linked to the characteristics of L2 form-meaning system, making rating linguistic complexity more objective and formal (Pallotti, 2015). The key features of a complex speech performance can be defined as the extent to which learners use rich, diverse, sizable, and elaborate lexis and structure (Housen & Kuiken, 2009). Given these features, four types of complexity can be distinguished, i.e. *morphological*, *phonological*, *syntactic* and *lexical*. Receiving far more attention, syntactic and lexical complexity are now two distinctive constructs that are operationalised through a set of different specific and general measures that tap into the breadth and depth of L2 structures regarding syntactic complexity, and density, diversity and sophistication regarding lexical complexity.

Moving to accuracy, L2 researchers argue that it is the most direct and explicit aspect of CALF (Housen et al., 2012). Accuracy is often associated with the degree of language correctness versus erroneous. Within this essence, Yuan and R. Ellis (2003) define accuracy as “the extent to which the language produced conforms to target language norms” (p. 2). This definition is consistent with that of Housen and Kuiken (2009), who define accuracy as the extent to which language performance diverges from a specific standard. It could be argued here that the disagreement between researchers on defining or identifying the standard norm, makes these proposed definitions problematic and controversial. Moreover, the calls to consider the gravity of errors, i.e. assessing learners' errors based on their effects on communication, as a more reliable feature of accuracy (Foster & Wigglesworth, 2016) is another issue that needs consensus among L2 researchers. Pallotti (2009) also advocates for considering the adequacy and acceptability of speech production as fundamental elements of accuracy. This can be true

in the sense that L2 learners may produce grammatically accurate utterances while they are still communicatively inadequate. Therefore, it is argued that any deviations from nativelike norms need to be considered as inaccurate or erroneous language performance.

Fluency is another construct not easy to define and operationalise for its multidimensional nature. Fluency can refer to a person's general language proficiency, characterized by perceptions of ease, articulateness, and smoothness of speech (Skehan, 1998). Housen and Kuiken (2009) point out that the challenge is to determine exactly which quantifiable linguistic phenomena contribute to perceptions of fluency in L2 speech. Lennon (2000) defines fluency more explicitly as "rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention under the temporal constraints of on-line processing" (p.26). This definition integrates many features of fluent speech (e.g., to speak in real-time without conscious planning, to be understood, and to formulate ideas into language), whereby fluency is seen mainly as the ability to speak spontaneously without relying on conscious planning. However, one can argue that L2 speakers with higher language proficiency can still speak fluently despite relying on conscious speech planning and processing. Moreover, considering accuracy in this definition as an aspect of fluency could be misleading for research within CALF framework which counts accuracy as a distinct construct. Given the manifold nature of fluency, Skehan (2003) classifies three sub-constructs of fluency that need to be captured, i.e. speed (speech rates and density), breakdown (pausing behaviour), and repair (dysfluency). Segalowitz (2010) confirms Skehan's fluency model by linking it to the quality of language performance in terms of the existence of automaticity, ease, speed and control, and the absence of dysfluency incidents and filled or silent pauses.

According to Skehan (2009), ideal speech performance within TBLT research needs to compromise "more advanced language, leading to *complexity*; a concern to avoid error, leading to higher *accuracy*; and a capacity to produce speech at normal rate and without interruption, resulting in greater *fluency*" (p.510). Based on this description, *complexity* is seen as dependent on learners' willingness to take risks by trying new forms, *accuracy* is seen as dependent on learners' focus on form, while *fluency* is achieved when learners prioritise meaning over form. Henceforth, research within TBLT context has devoted more efforts to investigate whether these aspects of L2 performance compete against each other or complete each other during the process of speech production. This bulk of research is motivated by the proposals offered by two leading theoretical frameworks, i.e. Skehan's (1998, 2014, 2015a) Limited Attentional

Capacity and Robinson's (2001, 2011a, 2015) Cognition Hypothesis which have informed research on TBLT generally and task complexity particularly.

2.6 Task Complexity

Task Complexity (TC) is a key issue in TBLT research that is often paired with Task Difficulty (TD) or used interchangeably by some researchers, whereas in fact each refers to a different construct (Bulté & Housen, 2012). From a cognitive approach to TBLT, TC can be defined as any attentional or information processing demands that are enforced on the task-performers by different degree of inherent task requirements (Robinson, 2001). TD is, however, associated with the task-performers' perceptions of the degree of difficulty of a task, based on their individual difference (Robinson, 2007). As will be discussed later in this section, the models of TC advocate different definitions for these two constructs, which poses more challenges on researchers in exploring each construct. The challenge that faces TC research and also faces this study is how to thoroughly conceptualise and systematically operationalise TC, not to mention how to clearly differentiate between TC and TD. One way to tackle this challenge is to understand TC not only from a narrow scope of TBLT and SLA but also from a broader scope of the neighbouring areas (e.g. psychology and management). This is assumed to advance our knowledge of how TC influences not only language performance but also human behaviour. Being the main focus of this study, TC will be carefully defined and distinguished from TD within TBLT context and other disciplines. TC will be then linked to the theoretical models of TC, i.e. the Limited Attentional Capacity and the Cognition Hypothesis. Finally, the diverse ways TC has been operationalised and investigated in task-based studies will be explained and discussed.

TC can be tackled from three broad standpoints, i.e. structuralist (structure of task), resource requirement (what resources a task recalls), and interactionist (performer-task interaction) (Liu & Li, 2012). Given a structuralist perspective, Wood (1986) proposes that TC is contingent on the given information, action required and final product of task. Building on this model, Wood identifies three aspects of TC: 1) component complexity (data and acts needed to perform a task), coordinative complexity (task input-output interaction), and dynamic complexity (i.e., changes in the input-output relationships). While the first two aspects of TC are internal due to task characteristics, the latter is external due to any changes that may take place regarding task conditions. Motivated by an information processing perspective, Bonner (1994) categorises TC

into three dimensions, i.e. input, processing, and output. Each element of this model depends on the quantity and quality of the offered information. Campbell (1988) argues that TC is highly dependent on the number of possible alternatives to do a task or number of possible outcomes of a task. In other words, the more options task takers have to perform a task or the more possible solutions a task has is assumed to make it more complex to accomplish. To sum up, TC from a structure-driven view can be seen as a result of manipulating task design with respect to amount and clarity of available information, kind and number of actions required, and the multiplicity of task outcomes, as well as the interaction between all aforementioned elements.

Moving to define TC from a resource requirement standpoint, TC “seems to be synonymous with task load or task demand” (Liu and Li 2012, p.556). Campbell (1988) links TC to the characteristics that “increase information load, diversity, or rate of change” (p. 43), which in turn tax task performers’ cognitive resources (e.g., reasoning, problem-solving, creative thinking). In this sense, increasing complexity of a task might impose higher demands on the cognitive, mental, memory, attentional, other information processing or even physical resources of task performers. Hence, these resources will be taxed based on what cognitive, memory, attentional, or physical efforts are required to successfully do the task. Consequently, manipulating TC can push task performers to devote the relevant resources with certain amounts needed during performance which can affect their behaviour and performance correspondingly.

TC from an interactionist stance, can be described as “a product of the interaction between task and task performer characteristics” (Liu and Lu, 2012, p.555). Thus, TC is a combination of the objective and subjective complexity. The latter is known also as task difficulty or perceived TC based on the performers’ individual variables (e.g., prior knowledge and skills). This leads to define TD from a broader scope as a distinctive construct to solve any confusion with TC. Though TC and TD may seem compatible, the two constructs signify different concepts according to the interactionists. Considering TC as a sub-component of TD is argued to be true as a difficult task is not always complex, whereas a complex task is likely to be difficult (Braarud & Kirwan, 2010). This is subject to a number of individual factors that can be anticipated prior to performing certain tasks through background questionnaires and/or interviews or identified after task performance through retrospective questionnaires and/or interviews. However, the risk entails in the reliability of task performers in rating TD objectively and identifying the factors that contribute to the perceptions of difficulty effectively.

Building on the various viewpoints of TC discussed above, Lui and Li (2012) suggest an inclusive definition for TC that can serve as a point of departure to operationalise TC in any discipline. They define TC as “the aggregation of any intrinsic task characteristic that influences the performance of a task” (p. 560). This definition emphasises that TC is multifaceted, task-dependent, and judged based on its impact on task performance and the performer’s behaviour and perception. Lui and Li also identify five key factors that contribute to TC which need to be considered in operationalising TC: 1) *output* (task outcomes or goals); 2) *input* (e.g., information cues, instructions); 3) *process* (e.g., steps, actions, tracks); 4) *time* (concurrency or pressure); and 5) *presentation* (mode or method). Researching TC within TBLT context has been largely influenced by this wide range of definitions and models of TC offered by other disciplines. Following are definitions of TC from TBLT research perspectives.

Within TBLT research, TC is defined as the amount of attention a task requires from learners to achieve its outcome (Skehan, 2001). It is also advocated that “tasks which are cognitively demanding in their content are likely to draw attentional resources away from language forms” (Skehan & Foster, 2001, p. 189). This viewpoint assumes that more cognitively demanding tasks are expected to consume L2 learners’ attentional resources, thus affecting their performance negatively. This definition can be further criticised for limiting the impact of TC to learners’ attentional resources only with a presumed negative impact. Robinson (2001) who proposes a more detailed definition, describes TC as “the result of the attentional, memory, reasoning, and other information processing demands imposed by the structure of the task on the language learner” (p. 29). This widely-adopted definition of TC appears more successful in highlighting the various components of TC. Based on this definition, Robinson relates TC to a combination of cognitive, attentional and other resource requirements arise from task design. Robinson’s definition further considers TC as task-dependent rather than being learner-dependent. Hence, Robinson distinguishes between TC and TD, as the latter is related to learners’ perceptions of task demands based on their individual differences (e.g., working memory, proficiency, aptitude) or affective factors (anxiety, motivation), while the former is linked with the inherent task demands imposed by task design.

R. Ellis (2003) argues that TC is the extent to which a task is inherently easy or difficult. Prabhu (1987) suggests that TC depends on the processed information of task in terms of its amount, tangibility, familiarity, and accuracy, in addition to the degree of reasoning required. Prabhu therefore claims that TC is a combination of cognitive and linguistic demands. So, manipulating

the information cues and the cognitive resources required in a task are the two ways to operationalise different degree of TC. Designing tasks with different levels of complexity is not only an important element of task design, but also a criterion to classify tasks within syllabus design based on increased complexity (Robinson, 2015). Complexity-driven syllabus design is assumed to help L2 learners achieve more balanced and simultaneous gains in all aspects of language performance (Skehan and Foster, 2005) or gradual L2 development due to the different functions of each level of complexity (Robinson, 2011a).

Regardless of whether TC is inherent or external to task, or whether it promotes performance and facilitates L2 development, or whether it is an effective principle to sequencing tasks, it is essential to consider its relationship with other variables and factors (e.g., attentional resources, cognitive resources, individual variables, task difficulty) that might play crucial roles in regulating the effects of TC. Thus, investigating TC from wider perspectives can offer a more comprehensive picture about the importance of TC in L2 performance and development, and hence identify robust criteria for sequencing tasks based on their degree of complexity. The Limited Attentional Capacity (Skehan, 1998, 2015a) and the Cognition Hypothesis (Robinson, 2001, 2015) have offered theoretical umbrellas for TC research in the last two decades to establish the relevant factors, task characteristics and task conditions that subsidise TC, and suggest sufficient criteria for sequencing tasks in terms of their complexity.

2.6.1 The Limited Attentional Capacity

Skehan's Limited Attentional Capacity (LAC) (1998, 2014, 2015a) model, known also as the Trade-off Hypothesis, proposes that learners' attentional resources are limited, and therefore a trade-off usually exists between specific dimensions of task performance (Skehan, 1998). Consequently, it is not always possible to promote complexity, accuracy, and fluency simultaneously in more complex tasks. According to LAC, the restricted amount of attention available during performing complex tasks force learners to decide whether to consume the free attentional resources to attend syntactic and lexical complexity or accuracy (Skehan, 2015a). Skehan claims that complexity-accuracy trade-off is due to the fact that the two aspects compete on the same pools of attention which explains why fluency does not compete with complexity or accuracy. According to LAC, the interaction between TC and L2 learners' limited attentional resources is assumed to result in tension between accuracy and complexity, and hence certain tasks foster fluency and accuracy, while other tasks advance fluency and complexity.

LAC has been theorised on the assumptions that attention and noticing are crucial for SLA (Schmidt, 1994), working memory is limited and a gateway to long-term memory (Miyake & Shah, 1999), attentional resources are limited and administer meaning-form competition (VanPatten, 1990), and it is possible to reallocate and assign these resources (Anderson, 1996) within the stages of Levelt's (1989) model of L1 speech production (Skehan, 2014). The tenets of LAC model have devoted notable efforts to explore what task characteristics and conditions push L2 learners to dedicate attention to specific areas and how to design tasks or improve their conditions in order to diminish the trade-off negative impact. One way to tackle this issue was to put forward a framework to manipulate and investigate TC more systematically which can advance understanding on how to sequence complex tasks to facilitate more enhanced language performance in all aspects. Hence, Skehan (1998) put forward his scheme of TC to guide research efforts in a more systematic way.

2.6.1.1 Skehan's (1998) Task Complexity framework

Skehan (1998) suggests the following scheme that can be employed to explore and classify TC: a) code complexity (language required); b) cognitive complexity (thinking required), which includes cognitive familiarity and cognitive processing; and c) communicative stress (task conditions). Skehan summarizes the factors that might influence TC as: a) number of participants or elements; b) abstractness of information; c) type of information; d) familiarity of task information; e) nature of operation required; and f) time pressure. Central to Skehan's framework is the assumption that attention and working memory are limited and that any improved performance in more complex tasks is dependent of the interaction between these limitations and the elements of TC. Given this framework, TC is reliant on what linguistic and cognitive resources a task requires, in addition to task conditions and information cues that may complexify, pressurise, ease, or focus task performance.

Skehan (2015a) argues that task characteristics and conditions can be manipulated to affect speech production differently at each stage of Levelt's model. Thus, Skehan (2014) classifies sets of variables that have a *complexifying* impact on performance at the Conceptualisation (e.g., abstract or unstructured information) versus variables with *easing* effects (e.g., concrete or familiar information). Another set of variables may lead to *pressuring* effects at the Formulation (e.g., time pressure, monologic tasks) versus variables that lead to *focussing* influences (e.g., online planning, dialogic tasks). Based on these categories, it is possible to attain combined complexity-accuracy enrichment by lessening the trade-off effects due to TC if a task, for

example, has a clear structure or familiar information, in addition to offering planning opportunities (Skehan, 2009).

Based on his framework, Skehan (2014) posits that tasks should be sequenced by selecting task characteristics that promote CAF at an appropriate level of TC in an attempt to channel learners' attention in a predictable way. However, it could still be argued that even tasks at an appropriate level of complexity will not always have that predictable influence on the processes of speech production. Therefore, it is sometimes not easy or feasible to control learners' attentional resources as Skehan predicts. Skehan's LAC model can be also criticised for underestimating the interaction between TC and learners' individual differences, affective factors, needs, and agendas while performing tasks. It is claimed elsewhere that the learner factor, i.e. what a learner brings to task, may result in guiding attention to unpredictable directions or extending the available attentional resources to attend more than one performance dimension. This imposes a need to identify more explicit and feasible variables that can be more predictable in directing learners' attention and other cognitive resources to advance certain aspects of performance. The LAC also underrates the positive effects of increasing TC on speech performance with the presumptions that learners have inextensible attention, unlike the Cognition Hypothesis which proposes different views on how L2 learners' attentional resources can be extended to overcome any trade-off effects as a result of any cognitive demands of TC, taking into account also the significant roles learners' variables can play.

2.6.2 The Cognition Hypothesis

Robinson's Cognition Hypothesis (CH) (2001, 2011a, 2015) advocates that learners' attentional capacity is not limited in the way LAC claims. Drawing on the work from cognitive psychology (e.g., Givón, 1998), Robinson states that human beings have multiple pools of attention which they draw on while engaging in more demanding tasks, and therefore learners can process more than one aspect of language simultaneously. Hence, the CH predicts that increasing TC can have a joint positive effect on speech performance in terms of complexity and accuracy at the expense of fluency which contradicts the speculations of LAC. While task conditions are more important for the LAC, the CH argues that task characteristics (e.g., resource-directing versus resource-dispersing variables), cognitive resources (e.g., working memory, aptitude, and reasoning) and affective factors (e.g., motivation and anxiety) play dynamic roles in regulating and promoting L2 performance (Robinson, 2007). Robinson argues that manipulating TC is assumed to promote L2 interaction, negotiation for meaning, noticing, and uptake. The CH also

claims that the individual variations between L2 learners cognitively and affectively have a greater impact on performance when tasks are more complex, since the high demands of more complex tasks activate learners' schemata and push them to bring their ability and affective factors into action (Robinson, 2015). Driven by a psycholinguistic perspective to SLA and deriving from neighbouring disciplines (e.g., cognitive psychology), Robinson has suggested a framework for researching and integrating TC into syllabus design. Building on the assumptions of his CH with regards attention, memory, learners' factors and task sequencing, Robinson's framework of TC attracts a considerable number of TBLT researchers and forms the main theoretical orientation to manipulate and operationalise TC.

2.6.2.1 Robinson's (2001) Task Complexity framework

Robinson (2001) has elaborated an operational taxonomy of task characteristics, i.e. the Triadic Componential Framework (TCF) to classify, sequence and research TC. As shown in Figure 3, Robinson (2001) classifies three categories of task demands implicated in real-world task performance: 1) task complexity (cognitive factors); 2) task conditions (interactive factors); and 3) task difficulty (learner factors), as criteria to classify, operationalise and sequence tasks based on their complexity. TCF considers the factors that affect complexity as 1) cognitive (resource-directing versus resource-depleting variables); 2) interactive (participation and participant variables); and 3) learner factors (affective and ability variables) (Robinson, 2011). Robinson (2005) posits that resource-directing variables (e.g., reasoning) create cognitive and conceptual demands. Thus, these variables direct learners' attention to certain aspects of language which are needed to meet the increased demands, resulting in promoting form, i.e. accuracy and complexity, but not fluency. (Robinson, 2005) further claims that the resource-directing factors have "the potential to direct learners' attentional and memory resources to the way the L2 structures and codes concepts, so leading to interlanguage development" (p. 4). For instance, increasing TC along this dimension through +reasoning by requesting learners to read other people's intentions and justify their actions will trigger the use of certain lexis (e.g., logical connectors and mental states verbs), resulting in advancement with respect to grammatical structures, lexis and accuracy. However, this opportunity is not available when performing tasks that require no reasoning.

Alternatively, increasing TC along the resource-dispersing variables (e.g., no planning time), creates performative and procedural demands, and thus imposes greater loads on working memory and depletes learners' attention to many non-linguistics areas of speech production,

affecting fluency, as well as accuracy and complexity in a negative way. LAC predicts the same effects of resource-dispersing variables, but it does not distinguish between the resource-directing and resource-dispersing dimensions in the same way as CH does. Furthermore, increasing TC by manipulating resource-dispersing variables “promotes greater control over, and faster access to the existing interlanguage systems of knowledge” (Robinson, 2011a, p.17). TCF proposes that resource-directing versus resource-dispersing variables interact and influence task performance in measurable and predictable ways (Robinson, 2010). Thereby, it is predicted that tasks that are complex along a resource-directing dimension (e.g., + reasoning) and simple along a resource-dispersing dimension (e.g., + planning) may lead to positive gains regarding all aspects of CALF. This is based on the assumption that per-task planning has positive consequences on lexis and fluency (Robinson, 2011a). However, not providing pre-task planning time would complexify the task along resource-directing (e.g., + reasoning) and resource-dispersing (e.g., - planning) aspects affecting both lexis and fluency negatively.

| <i>Task Complexity (Cognitive factors)</i> | <i>Task Condition (Interactive factors)</i> | <i>Task Difficulty (Learner factors)</i> |
|--|---|--|
| (Classification criteria: cognitive demands) (Classification procedure: information-theoretic analyses) | (Classification criteria: interactional demands) (Classification procedure: behavior-descriptive analyses) | (Classification criteria: ability requirements) (Classification procedure: ability assessment analyses) |
| <i>(a) Resource-directing variables making cognitive/conceptual demands</i> | <i>(a) Participation variables making interactional demands</i> | <i>(a) Ability variables and task-relevant resource differentials</i> |
| +/- here and now | +/- open solution | h/l working memory |
| +/- few elements | +/- one-way flow | h/l reasoning |
| -/+ spatial reasoning | +/- convergent solution | h/l task-switching |
| -/+ causal reasoning | +/- few participants | h/l aptitude |
| -/+ intentional reasoning | +/- few contributions needed | h/l field independence |
| -/+ perspective-taking | +/- negotiation not needed | h/l mind/intention-reading |
| <i>(b) Resource-dispersing variables making performative/procedural demands</i> | <i>(b) Participant variables making interactant demands</i> | <i>(b) Affective variables and task-relevant state-trait differentials</i> |
| +/- planning time | +/- same proficiency | h/l openness to experience |
| +/- single task | +/- same gender | h/l control of emotion |
| +/- task structure | +/- familiar | h/l task motivation |
| +/- few steps | +/- shared content knowledge | h/l processing anxiety |
| +/- independency of steps | +/- equal status and role | h/l willingness to communicate |
| +/- prior knowledge | +/- shared cultural knowledge | h/l self-efficacy |

Figure 3. Robinson's TCF for task classification (Robinson, 2015, p.96)

As the first group of TC variables lead to inherent cognitive demands, the second set, i.e. Task Conditions can lead to interactional demands regarding ‘*participation dimension*’ and interlocutor demands with respect to ‘*participant dimension*’. The CH assumes that a task

condition that requires learners to transmit information in one direction versus two directions or come up with one closed solution versus negotiate for various solutions will bring different interactional demands into play. On the other hand, the similarities or differences between participants in terms of their proficiency, gender, background or other variables impose certain effects on those participants, and thus influence their language performance (Robinson, 2015).

Task Difficulty is the third set of factors which comprises learners' ability and affective variables. The CH makes predictions about potential moderation between what learners bring into tasks and the inherent TC demands (Robinson, 2015). It is therefore assumed that variations in learners' ability factors (e.g., working memory, reasoning, aptitude) and affective factors (e.g., motivation, anxiety) will have a major impact on learners' language performance and perceptions of TD. According to the CH, learner factors can serve as parameters that regulate learners' perceptions of TD, and therefore they should be considered as criteria in designing tasks (ibid). Learner factors are assumed to determine how difficult the same task is for each participant based on his/her skills and traits. This is why, it is important for researchers to gather data prior and post tasks about the participants' individual differences which can quantitatively and/or qualitatively describe between-participants variations in task performance as well as detect whether learner factors mediate TC effects.

Motivated by the claims that pedagogic tasks with increased cognitive TC foster SLA (Long, 1985; Merrill, 2006; Prabhu, 1987; Samuda & Bygate, 2008), Robinson (2011a) posits that tasks need to be sequenced based on increased complexity in a way that resembles the demands of real-world target tasks. Therefore, it could be deduced that the main pedagogic goal of the CH is to offer a rationale for sequencing tasks in a way that leads to the promotion of L2 performance and learning. Hence, TC should serve as the theoretical basis for designing and sequencing tasks within syllabus design and implementation. Robinson (2015) mentions several pedagogical functions for having a continuum of tasks sequenced from simple to complex. Robinson claims that each level of TC assists learners to advance their IL in a different way. For example, simple tasks mitigate the effects of the shallow IL and boost the current knowledge. The gradual increase in TC enables L2 learners to faster access and automatize new forms and structures. As tasks become more complex, learners are pushed to extend their IL and take risk to recycle more advanced and newer language. In a long run, old language will be stabilised and new language will be attempted, resulting in sustained IL development (ibid).

To guide researchers, teachers, and syllabus designers and make sequencing tasks more feasible, Robinson (2010) suggests a continuum of TC. As shown in Table 1 below, the SSARC is a four-step model for sequencing pedagogic tasks based on gradual complexity. Simple tasks along both resource-directing and resource-dispersing variables go first to simplify and stabilise learners' IL. Increasing complexity along dispersing variables only in the second step is meant to support the speed access and automatization of already learned linguistic items. The third step comes next to help learners restructure and stretch the acquired language by only increasing the demands along resource-directing variables. The final option of TC scale is making the task more complex at both sides to encourage learners to complexify their IL and attempt new forms.

Table 1. Robinson's (2010) SSARC Model for increasing TC.

| TC scale | Function | Abbr. | Resource-directing | Resource-dispersing |
|----------------------|------------------------|--------------|---------------------------|----------------------------|
| 1. simple | Simplify and Stabilise | SS | simple | simple |
| 2. complex | Automatise | A | simple | complex |
| 3. + complex | Restructure | R | complex | simple |
| 4. ++ complex | Complexify | C | complex | complex |

To sum up, the triad dimensions discussed earlier, i.e. *TC*, *Interactive Factors* and *TD* are the key elements in: 1) integrating and manipulating TC in task design; 2) classifying and sequencing pedagogic tasks based on increased complexity; and 3) operationalising and researching TC. As mentioned before, TC variables contribute to the amount of intrinsic cognitive demands of a task, and hence inform decisions on designing and sequencing tasks. The interactive factors influence the type and amount of language performance, and hence inform decisions on task conditions in terms of participation and participants. TD variables contribute to explain between-participants variations in performing the same task, and hence inform the validity of TC manipulation and the combined TC-TD effects. Since TD variables are out of reach and cannot be operationalised or manipulated, researchers can use the data obtained from learner ability and affective factors to judge how effective TC operationalisation is, how successful TC sequencing is within syllabus design, and how the individual variables interact with TC to affect speech performance. However, the aforementioned issues emerging from TC frameworks are still debatable and subject to disagreement amongst the tenets of the LAC model and the CH.

2.6.3 Skehan's LAC versus Robinson's CH

Given the feasible TC frameworks and schemes they suggest, it is no wonder that the LAC and the CH models form the cornerstone of TC research and motivate its research agendas. However, building on different theoretical perspectives and empirical evidence, the two models disagree on certain issues concerning TC. Both models have different views with regards to TC effects on performance and development, multiplicity of attentional resources, TC sequencing, TC manipulation, learner factors, and TC-TD distinction and interaction.

Whilst the CH deems increasing cognitive TC as beneficial to L2 performance and development, LAC claims that TC at inappropriate level can cause problems for L2 learners due to limitations in their attention, working memory and language proficiency. Additionally, the CH claims that resource-directing variables not only enhance language performance but also have positive consequences on acquisition. The LAC, in contrast, does not posit similar claims expect for post-task activities (e.g., transcribing task performance by the participants), which are argued to facilitate acquisition (Skehan, 2014). However, this incompatibility between the two models in terms of the predicted impact of TC on performance and acquisition is a domino effect of the way each model understands the role of attention, task sequencing and learner factors.

As its name suggests, the LAC model argues that learners' attention and working memory are of limited capacity, leading to trade-off effects during performing complex tasks. As a result, learners are likely to sacrifice one aspect to attend the other (complexity or accuracy), which means that either complexity or accuracy will suffer. The LAC gives more prominence to the psycholinguistic processes underlying L2 speech production which can be best explained and understood through Levelt's monolingual model. LAC, therefore, attributes L2 learners' trade-off decisions to the different consequences of lack of attention and memory capacity on each stage of speech production. The CH, however, believes in the multiplicity of learners' attentional resources and that learners have the ability to stretch their attention when they are forced to do so. Though, the CH does not explain how learners extend their attention, this flexibility is chiefly advantageous during monologic tasks, initiating the availability to attend to all aspects of performance simultaneously if attention extension is facilitated (e.g., through planning). Therefore, it is predictable that performing more complex tasks will not be costly regarding complexity and accuracy as the LAC assumes, but conversely, TC is expected to bring positive gains for the two aspects concurrently. However, both the LAC and CH meet on their predictions about the negative consequences of TC demands on fluency.

Regarding task sequencing, this issue is central to the CH which proposes schemes of tasks based on increased TC to inform decisions regarding task sequencing at the levels of syllabi and classroom. As a guidance for syllabus designers, practitioners and researchers, the CH recommends the SSARC model as a plan that serves functional and pedagogical options with regards to a step-by-step increase of TC. On the other hand, this issue is not a prime focus for LAC, and hence to the research agenda of its tenets. As for the classification of TC variables, the CH is built on the division between two categories of TC variables, i.e. resource-directing versus resource-dispersing which generate different effects on performance and diverse consequences on L2 development. These variables are recommended as feasible methods to operationalise and research TC with predicted consequences on language performance. Nevertheless, the LAC does not make this classification of variables that drive attention in different directions causing positive or negative influence on each aspect of learners' language performance.

As regards learner factors, the CH argues that variations between learners in their ability and affective variables interact with TC demands interchangeably. The CH anticipates that learners' ability factors, which are assumed to be stable, and their affective factors which are assumed to be unstable, to influence how learners approach a task and engage in language performance. Consequently, they are predicted to influence the outcomes of task performance and the perceptions of TD level (Robinson, 2011a). Meanwhile, the LAC agrees with the CH on the key role of the individual difference between learners when performing complex task. However, the LAC does not consider learner factors as a separate dimension that is associated with TD.

The CH, further, distinguishes between TC (task characteristics) and TD (learner factor), whereas the LAC does not propose for any division or interaction between TC and TD, and does not consider TD as an element of TC. While the CH defines TD as the extent to which a task is easy or difficult based on learners' rating, the LAC links TD to the amount of attention required to perform a task and sees it as "inherent in the task, rather than learner-dependent" (Skehan, 2014, p.6). However, the disagreement between the two TC models has a positive impact on TC research as it motivates researchers to test the predictions of each model to obtain more robust empirical evidence to back one of the two models. This track of research has operationalised TC differently based on the theoretical framework it follows. Following is a discussion on how TC is operationalised in task-based research and how TC will be operationalised in this study.

2.7 Research on Task Complexity

Empirical studies that examine TC as a key construct in task design aim at exploring its impact on: 1) language performance measured by CALF; 2) L2 development and acquisition; 3) learners' perceptions of TD; 4) mediation of individual difference; 5) interaction, negotiation and uptake; 6) meaning-form attention; 7) automaticity; and 8) sequencing decisions. Furthermore, research on TC attempts to identify what task features, conditions and factors affect or mediate the effect of TC. Owing to the fact that the CH has proposed a more feasible framework to manipulate TC, the focus here will be on how previous research has operationalised resource-directing and resource-dispersing variables with a focus on studies that were interested in TC effects on CALF.

Looking back at Robinson's (2010) TCF and previous TC research, it could be observed that some variables have attracted more attention than others. In terms of resource-directing dimension, TC has been chiefly operationalised through *here-and-now*, *number of elements*, and *reasoning*. Regarding resource-dispersing dimension, TC research has largely considered manipulating *planning time*, *task structure*, *number of tasks*, *number of steps* and *prior knowledge*. Researchers use -/+ to indicate the absence or the presence of TC variables in each condition, or ++ if more than one level of TC is employed. It is worth mentioning that a separate section (see Section 2.8) will be dedicated to how intentional reasoning (the focus of this study) has been operationalised in previous research. Some studies manipulate TC using only one resource-directing or resource-dispersing variable, whereas other studies manipulate the two dimensions simultaneously or sometimes use two or more variables from each dimension. The CH predicts positive effects of the resource-directing variables on complexity and accuracy but not fluency, whereas increasing TC further along resource-dispersing variables affects all aspects of CALF negatively.

2.7.1 Research on resource-dispersing variables

Starting with resource-dispersing variables, manipulating planning time has attracted much more attention than any other variables. Two types of planning time can be acknowledged in task-based studies, i.e. pre-task and within-task (R. Ellis, 2005). Pre-task planning can be manipulated by giving learners time to plan ahead what they want to say (strategic planning), or giving them opportunities to repeat the whole task ahead of the real performance (rehearsal). Strategic planning can be controlled by increasing, reducing or removing the time allocated for

planning, whereas rehearsal can be manipulated by the number of repetitions learners are allowed to do (ibid). Regarding within-task planning (also known as online planning), it involves allowing time for learners to prepare while engaged in task performance by controlling time pressure. The CH predicts that offering planning time is expected to assist L2 learners to faster access and retrieve the required linguistic components resulting in promoting complexity, accuracy and mainly fluency with positive consequences on automaticity (Robinson, 2015). Therefore, increasing TC along a resource-directing variable and keeping it simple along a resource-dispersing dimension by offering opportunities for planning time would enhance all aspects of performance, i.e. CALF. Similarly, the LAC presumes that due to the limitation of attention, not offering planning time during complex task would result in consuming the available attentional resources, and therefore fluency will suffer, as well as the other aspects, whereas offering planning time will facilitate freeing more attentional and memory resources to attend both meaning and form with positive consequences on CALF. While the LAC does not suggest that planning can mitigate the trade-off effects when task is more complex, both models meet regarding the impact of the presence or absence of planning on task performance.

Task structure is the next dominant variable to be tackled in TC research. *Task structure* is associated with the extent to which task sequence is clear or logical in terms of information, timeline or problem-solving elements (Skehan and Foster, 1999). Consequently, task structure can be manipulated by controlling the logical sequences of task steps, familiarity of the given information, or the presence of problem-solving components. This means that narrative tasks with loose or disconnected events or unclear timeline or the absence of some problem-solving elements will be more cognitively complex than structured counterparts (Tavakoli and Foster, 2011). According to the CH, unstructured tasks deplete attention from attending their linguistic requirements resulting in negative effects on all aspect of performance, while structured tasks release more attention to help speeding up access to the mental lexicon, and hence fluency improves. However, the LAC argues that structured tasks that entail clear and tight storylines or clear problem-solving components free up more attentional resources resulting in advancing accuracy and fluency, while unstructured tasks will advance complexity at the expense of accuracy and fluency (Skehan and Foster, 1999).

Other TC variables under resource-dispersing category include +/- *single task*, which requires learners to focus only on one task in the simple version but includes a secondary one in the more complex version. The CH operationalises +/- *single task*, by requiring participants to give

directions using a map with marked routes in the simple task, whereas the complex version uses a map with unmarked routes (Robinson, 2001). Hence, learners in the complex task will need to do a dual-task, i.e. thinking about the routes while giving directions. Other operationalisation of dual-task condition is to ask participants to do a secondary task (e.g., visual, auditory or physical) that taps into different pools of attention. For example, Declerck and Kormos (2012) operationalised *-single task* by asking the participants to do finger-tapping on a keyboard while doing an oral narrative task. Such dual task demands are assumed to tax the learners' ability to access the language required or monitor the language produced resulting in adverse effects on CALF (ibid).

Another variable with depleting effects on attention is *+/- few steps* which can be operationalised through manipulating the number of steps required to do a task. A simple version requires a few steps, whereas a complex version needs many more steps. Finally, *+/- prior knowledge* as a resource-dispersing variable is operationalised by the extent to which learners are familiar with the content of a task. For example, Robinson (2001) has manipulated *+/- prior knowledge* as being familiar (+prior knowledge) or unfamiliar (-prior knowledge) with a map route. As mentioned above, the presence of all resource-dispersing variables (+) will make the task simple resulting in strengthening learners processing in terms of language retrieval and self-monitoring with positive gains on all aspects of language performance. On the other hand, the absence of those variables (-) will deplete learners' attention over non-linguistic demands, tax attention and memory, and create performative and procedural burdens resulting in opposing effects on CALF.

2.7.2 Research on resource-directing variables

Moving to operationalising the resource-directing variables of TC, *+/- Here and Now*, *+/- few elements and -/+ reasoning* have been widely investigated as factors that contribute to task-inherent cognitive demands. The *(+/-) Here and Now* embraces two versions of task design, i.e. a simple version (Here-and-Now) which requires learners to talk or write about what is happening at the moment with the presence of a visual support versus a complex version (There-and-Then) which requires addressing events that happened in different time and place with the absence of visuals (Robinson, 2011a). It is anticipated that the later condition will pose more cognitive demands at the Conceptualisation and will tax memory because the events happened elsewhere which requires more efforts to visualise them. Thus, performing a task under There-and-Then condition will direct speakers to refer to past tense morphology compared to referring

to progressive aspect of the events in the Here-and-Now. The CH predicts that increasing TC along There-and-Then will orient “learners’ attentional and memory resources to the way the L2 structures and code concepts” (Robinson, 2011a, p15) are represented in L2. As a result, complexity and accuracy are expected to gain benefits at the expense of fluency.

While some studies examine the effects of manipulating TC along +/-*Here-and-Now* factor on L2 development in terms of past tense (e.g., Révész, 2009) or self-repairs behaviour (e.g., Gilabert, 2007a), more studies have been interested in its effect on L2 language performance measured by CALF. Employing monologic narrative tasks, Gilabert (2005) and Robinson (1995) found that increasing complexity along *-Here-and-Now*, was associated with improved accuracy at the expense of fluency, whereas mixed results were obtained with regards to complexity. Malicka and Sasayama (2017) confirmed in their meta-analysis on TC that the complex task, i.e. *-Here-and-Now* was only associated with improved syntactic complexity.

Manipulating TC along +/- *few elements*, involves identifying a few distinctive elements in the simple version which are easier to notice versus many similar elements in the complex counterpart (e.g., many people in a place, many routes in a map) (Robinson, 2001). It is expected that doing a task under *-few elements* condition will require participants to use wider range of linguistic components to establish relations or distinguish between many similar elements. Therefore, more conjunctions, subordinations, adverbs, adjectives and formulaic sequences will be needed. Operationalising +/-*few elements* by increasing the number of places in a map, Robinson (2001) found that the complex task produced higher accuracy and lexical variety but reduced fluency as he predicted. Furthermore, Michel (2011) in argumentative tasks about dating, and Levkina and Gilabert (2012) in advice-giving tasks about holiday destinations found also that the complex version with many elements elicited increased lexical complexity and decreased fluency. Thus, the studies reported above suggested a clear pattern in terms of the positive effects of TC manipulated along +/-*few elements* on lexical complexity and negative effects on fluency, whereas inconsistent patterns were found with respect to syntactic complexity and accuracy.

2.7.3 Research on reasoning as a TC variable

Manipulating TC along *-/+reasoning* as a resource-directing variable reveals a more inconsistent picture regarding the effects of increasing reasoning demands on L2 learners’ speech performance. This can be attributed to the disagreement among researchers on how to

define and operationalise each type of reasoning in a way that fits TC research agenda. A scarce number of studies focused on the effects of reasoning on CALF, while other studies combined reasoning with other TC variables. Under the resource-directing dimension, Robinson (2011a) identifies three types of reasoning demands that can be operationalised to increase TC, i.e. a) *spatial reasoning*, which involves navigating places and giving directions; b) *causal reasoning*, which involves explaining why events have happened; and (c) *intentional reasoning*, which requires understanding and explaining the rationale beyond other people's actions. The latter type of reasoning will be subject to a thorough discussion in the next section.

Operationalising different resource-directing variables simultaneously including reasoning, Gilabert (2007a) employed a monologic decision-making task based on a fire-rescue operation to investigate the effects of increasing reasoning demands on self-correction behaviour as a measure of accuracy. The complex version, i.e. *+reasoning* required the 42 participants (low and high proficiency) to set priorities, make a series of decisions and justify their decisions to rescue a number people from a burning building. The situation in the *+reasoning* task was more complicated and dynamic than in the *-reasoning* version. Gilabert found that the reasoning demands did not influence self-repairs behaviour, and thus the hypothesis was not confirmed. Failing to initiate more repairs in the complex task could be due to the participants' language proficiency. This was confirmed through the low-proficiency learners who initiated less repairs than the high-proficiency learners which could be due to their shallow linguistic knowledge. However, the way Gilabert operationalised reasoning might involve two types reasoning, i.e. spatial (navigating the building) and casual (justify the decisions) which might trigger more reasoning demands even in the simple task.

Baralt (2013) investigated how TC manipulated by *-/+* intentional reasoning promoted L2 development through recasts. The 84 participants who were L2 learners of Spanish received recasts on the correct use of Spanish past subjunctive. A pre-test and two post-tests were employed to measure the participants use of subjunctives. The data were collected over a whole year based on one-to-one interactive tasks with the researcher. The tasks were performed through face-to-face (FTF) versus computer-mediated communication (CMC) interaction. Using dialogic story retell, each participant had to collaborate with the researcher to retell two stories in the past tense based on a set of pictures. The reasoning tasks involved retelling the story and explaining the characters' intentions, whereas in the *-reasoning* version, the participants were provided with prompts that unfolded the characters' intentions, and therefore

they were required to only tell and describe. The results revealed that in the FTF mode, performing the complex task led to more L2 development in terms of use of subjunctives, unlike in CMC mode. The results indicated that corrective feedback through recasts worked more effectively when more complex tasks were performed FTF. The findings which showed that the mode of the tasks moderated the effects of TC of this study highlighted the importance of the interaction between task modality and TC in advancing language development.

Malicka (2014) investigated the combined effects of two variables of TC, i.e. *-/+reasoning* and *+/-few elements*. Three levels of TC were employed, i.e. simple, complex and +complex. The tasks were three monologic problem-solving situations about reallocating hotel guests. The *+/-few elements* variable was operationalised through manipulating the number of rooms features that would suit each guest, whereas *-/+reasoning* was operationalised through increasing the number of cognitive operations required to complete the tasks effectively which might include endorsing, regretting or justifying decisions. The study confirmed that the complex versions elicited higher accuracy and lexical complexity, but lower fluency. However, mixed results were obtained in terms of syntactic complexity. These results lent support to the predictions of the CH except for syntactic complexity. This study pinpointed the need to adopt a continuum of TC manipulation rather than dichotomous models which could advance our understanding about the ceiling effects of TC as Malicka suggested.

Reviewing a number of studies that manipulated TC along resource-directing and/or resource-dispersing variables depicted a vague picture about the effects of TC on L2 performance and development. While some studies recruited binary levels of TC, other studies investigated a continuum of increased TC. Driven by the predictions of the CH that increasing TC along resource-directing variables would advance complexity and accuracy at the cost of fluency and that increasing TC along resource-dispersing variables would result in reducing language performance in terms of CALF, only few studies provided full support to the predictions of the CH. A recent analysis of previous TC studies by Jackson and Suethanapornkul (2013) confirmed that TC had positive but negligible effects on accuracy and lexical complexity, whereas the effects on syntactic complexity and fluency were negative and small. These findings backed the CH predictions with regards to all aspects except for syntactic complexity. However, the most recent meta-analysis on TC research by Malicka and Sasayama (2017) offered more support to the CH predictions in terms of the resource-dispersing dimension but not for the resource-directing dimension with only one exception, i.e. *+/-here-and-now*.

Regarding *-/+reasoning demands*, Malicka and Sasayama revealed that the complex tasks which required reasoning demands had positive effects on accuracy and lexical complexity and a negligible negative effect on fluency, whereas syntactic complexity remained unaffected by increasing reasoning demands. This can be alarming for the accumulated research that employs the Cognition Hypothesis as the theoretical reference, and particularly, studies that examine the effects of manipulating TC along reasoning demands. These mixed results bring more attention to the inconsistent and unsystematic conceptualisation and operationalisation of reasoning as a factor to increase TC. However, this study which employs intentional reasoning to manipulate TC will attempt to address this issue and make novel contributions by re-defining and re-operationalising intentional reasoning.

2.8 Intentional reasoning

Intentional reasoning (IR) is one of three types of reasoning demands branded by the CH under resource-directing variables, i.e. *spatial*, *casual* and *intentional*. While spatial reasoning is associated with navigation of places, and causal reasoning involves clarification of causes, IR requires explanation of people's intentions. To guide TC researchers in their operationalisation of the construct, Robinson (2007) defines IR as “understanding and explaining the motives, beliefs and thoughts which cause others to perform certain actions” (p.194). Building on this definition, the CH anticipates IR requirements to orient L2 learners’ attentional resources to use more complex syntactic structures and advance lexis to articulate the demands of IR. As a result, performing complex tasks that require IR will amplify a focus on form and push L2 learners to produce language that is characterised with higher syntactic complexity, lexical complexity and accuracy at the expense of fluency (ibid).

Though Robinson’s definition has been widely adopted in IR studies, it fails to grasp the magnitude of the other elements that IR involves. IR as a complex construct may implicate more than just explaining motives or thoughts as Robinson argues. IR demands may require learners to read the thoughts of others, explain their desires, predict their actions, justify their decision-making and draw true conclusions about the consequences of their intentions (Bratman, 1987). This complicated process, as will be discussed in this section in more detail, is assumed to pose substantial burdens on learners’ various resources including attention and working memory with a direct impact on the characteristics of the language used to elucidate IR requirements.

Compared to previous research that investigates the resource-directing factors which contribute to TC (e.g. +/- *Here and Now*, +/- *few elements*), studies exploring IR as a variable to increase TC are scarce in number, unsystematic in terms of conceptualisation and operationalisation of IR, and inconsistent regarding their findings. Consequently, there is a persistent need to borrow from the neighbouring disciplines to fine-tune the current definitions of IR before any operationalisation of this construct is proposed. This study is therefore designed to address this research gap by redefining and re-operationalising IR in a more systematic and feasible approach.

Building up on concepts in cognitive psychology, Leighton (2004) defines IR as the process of drawing conclusions about people's intentions which can be linked to strategies of problem-solving or decision-making which others employ. These conclusions can be generated deductively or inductively based on a number of available grounds, and that the accuracy of these conclusions are highly dependent on the accuracy of the available premises (ibid). Bratman (1987) defines intentional reasoning as a state of mind, which allows us to characterize actions as done intentionally, or with some intention, or unintentionally. Therefore, the main components of intentionality are argued to be beliefs, desires, planning and coordination. Whenever someone intends to do something, this act will be intended to satisfy a current desire, which in turn, needs to be consistent with one's beliefs to guarantee effective planning and successful execution of the intended act.

According to Bratman (1987), IR combines both *intentions*, which are the acts one intends to do to satisfy a desire and *predictions* which are the unintentional consequences of executing that intention. For example, in the video story that has been employed in this study in the +reasoning task, the characters intend to make their car function as a plane to satisfy their desire to fly. In coherence with their belief that they can fly, they have tried several tricks to satisfy the practical instrumental requirements to make their car able to fly. Therefore, one thing they do get rid of the engine to make the car lighter in order to fly. However, a conclusion they have failed to predict is that their car will not start without an engine. Hence, some of these conclusions are intentions (make the car fly), while some are predictions (the car will not start without the engine). The characters could see the first conclusion which is coherent with their desires and intentions to fly, i.e. make their car lighter to fly. Nevertheless, they have failed to foresee the second conclusion because it is not coherent with their intention to fly, i.e. losing their car engine. Therefore, the characters in this story have failed to "evaluate these scenarios

as wholes and thereby arrive at a complex intention” (Bratman, 1987, p. 144). In other words, reasoning about other people’s intentions involves meeting two level of complex IR, 1) reading people’s thoughts, desires, motives and intentions to perform a certain action, and 2) predicting the unintentional consequences of that action. Building on Robinsons’ definition of IR, besides a wider scope of cognitive psychology, this study redefines IR as ‘reading other people’s thoughts and understanding their desires, beliefs and motives to draw true conclusions about what they intend to do, why and what consequences follow.’ For the purpose of this study, this broader definition of IR will serve as the base to propose a more systematic framework to operationalise IR in responding to the need to investigate this construct more thoroughly.

It can be concluded from a cognitive perspective that IR can comprise four successive steps, 1) observing actions; 2) understanding beliefs and desires; 3) drawing conclusions; and 4) predicting actions and their consequences. IR therefore involves foretelling peoples' actions and reactions based on their abilities, views, desires and knowledge (Astington & Baird, 2005). Meeting these requirements, reasoning about intentionality is estimated to pose higher attentional, memory and cognitive demands on L2 speakers because this process is assumed to be sequential (Leighton, 2004) and dependent on a series of premises and steps, in addition to the coordination between these steps (Gilhooly, 2004). Therefore, the more premises and steps this process involves and the less coherence these steps are, the more resources will be taxed and consumed to attend these IR demands. This can lead to conclude that performing language tasks that require both describing actions and explaining the intentionality behind them is assumed to be more challenging and cognitively demanding than performing tasks that only require describing actions.

From a psycholinguistic perspective, IR requirements as explained above will cause additional load at the Conceptualisation stage (Levelt, 1989) during planning and preparing for the preverbal messages needed to articulate the intentions and thoughts of others, justify their intended actions and predict any unintentional conclusions, in addition to linking between the various components of IR. When L2 learners are induced by task instructions and content to attend and meet IR requirements as a task outcome, they get engaged in explaining intentionality, and hence language of certain characteristics will be at stake. Therefore, I argue here that operationalising IR carefully at the level of task instructions, as well as task content is important to guarantee that each participant is aware of what each task requires in terms of intentionality. This two-level operationalisation of IR has the potentials to clarify the IR

requirements prior to the task, encourage engaging the participants in reasoning, trigger the use of more resources to meet these demands, and thus influence the characteristics of learners' speech performance.

As Robinson (2007) argues, a task that requires IR will encourage L2 learners to adopt linguistic structures with higher complexity (e.g., subordinating conjunctions) to create cohesion between intentions, actions and predictions. Hence, learners will be probably oriented to use lexical items with higher complexity (e.g., mental states verbs, adverbs of uncertainty). It is therefore expected that the characteristics of the language required to express intentionality as described above will also facilitate the use of more formulaic sequences (e.g., I think, it could be, it seems that). Such emphasis on expending more complex and formulaic language will have positive consequences on accuracy but may lead to reduce fluency as a result of such focus on form. As IR demands require speakers to maintain decision-making and process cognitively and linguistically at higher capacity during performance, more pauses, hesitations and repairs are the expected outcome of this process, causing decreased fluency. This is also in line with the CH predictions that IR demands have the effects to raise L2 learners' language in terms of complexity and accuracy but not fluency. Following previous research, one aim of this study is to test these predictions using more robust methodology and attend the criticism the previous studies have received.

The two studies that examined the impact of increasing TC along IR variable on oral speech performance measured by CALF, i.e. (Ishikawa, 2008; Robinson, 2007), showed inconsistent operationalisations of IR and also different results. Robinson (2007) used a continuum of TC through three collaborative tasks based on picture strips which were performed in dyads by 42 Japanese students of English. Each task involved a story-narrator who was required to sequence the pictures and tell a story to a listener, who was requested to rearrange the pictures based on the narrated story and initiate clarification requests and confirmation checks. Requiring the story-teller to do two tasks, i.e. sequencing and speaking, is assumed to tackle *+dual task* as a resource-dispersing variable. The narratives were one-way and closed tasks in terms of task conditions and participation dimensions. IR was operationalised through the requirements to understand and explain the intentions of the characters in the stories. The simple task entailed explaining the intentionality of only one character, whereas the more complex tasks involved more characters and their intentions were reliant on others' ideas and desires. Robinson employed both global and specific measures to operationalise speech performance.

Though IR demands resulted in more complex performance measured by specific measures (e.g. number of psychological and cognitive state terms), the complex tasks did not generate more complex and accurate language or even reduced fluency as the CH predicted using global measures. Moreover, lexical complexity as measured by type-token ratio was significantly higher in the simple task which pointed in the opposite direction of the hypothesis. These contradictory findings could be attributed to the way Robinson operationalised IR or to the differences between the three tasks in terms of IR requirements. Furthermore, designing tasks that are also complex along the resource-dispersing dimension, i.e. +dual task might have contributed to these results as it could be argued that the effects of IR confounded with the effects of the dual task. The clarification and confirmation checks initiated by the listener could be another factor that diminished or moderated the effects of IR on speech production. It was therefore not feasible to claim that the variation in the participants' performance was solely due to the IR demands. Another explanation for these results could be regarded to the failure of task instructions to encourage the participants to sufficiently attend reasoning while speaking.

Ishikawa (2008) operationalised IR by requiring the participants to reason about modifications in human relationship based on job mistakes at workplace. The participants (N=24) who belonged to different proficiency levels (low to upper-intermediate) performed three monologic tasks which required them to report to their boss about hypothetical changes in the relationships between staff based on several trouble triggers. Three levels of IR were employed, i.e. no reasoning, simple reasoning, and complex reasoning. IR demands were increased by manipulating the number of staff members and number of job mistakes. Three measures were used to operationalise fluency and one measure for syntactic complexity, lexical complexity and accuracy. The findings of Ishikawa (2008) showed that the more complex tasks with IR demands produced higher complexity measured by *S-nodes per T-unit* and *Guiraud 2000*, and higher accuracy measured by *percentage of error-free T-units*, whereas fluency decreased as measured by *pruned and unpruned speech rates*, and *number of repairs*. The findings were in line with Ishikawa's predictions and lent full support to the anticipations of the CH which advocated for joint development in terms of complexity and accuracy at the expense of fluency.

The study design and IR operationalisation of Ishikawa could be still questionable, despite gaining predictable results. The control task, i.e. no-reasoning condition could be challenged following the line of research discussed earlier, which did not suggest IR as a dichotomous construct. Instead, IR should be operationalised as a continuum variable with less or more

reasoning demands but not no IR demands. One could argue that there was no guarantee that the participants in Ishikawa's study would not reason in the no-reasoning condition. In terms of IR operationalisation, reporting about troubles in relationships between people could be problematic because it did not tackle all elements and levels of IR. Hence, Ishikawa's operationalisation of IR failed to reflect the multifaceted nature of reasoning. Furthermore, the lengthy task instructions and the offered planning time might have helped the participants in meeting the high demands of IR in the complex tasks, and thus affected the findings of this study. Finally, Ishikawa did not control for the effects of variations between the participants in terms of their language proficiency which might have mediated the effects of IR.

Reviewing TC literature and particularly studies investigating IR highlights problematic issues that need to be addressed and research gaps that need to be filled. It can be concluded that the decisions researchers have made with regards to operationalisation of reasoning, selection of tasks and designation of task instructions have contributed in the contradictory findings. The need to carefully re-defining and re-operationalising the constructs under investigation is another lesson that has been learnt from the literature review, not to mention the key role task instructions can play in directing the participants' attention and other cognitive resources towards certain aspects of performance.

As this study is mainly interested in investigating the effects of IR demands on L2 language performance and perceptions of TD, it will manipulate IR solely as a resource-directing variable, and will avoid any simultaneous manipulations of other directing-dispersing variables to eliminate any confounding results, as reported in some of the previous studies. Therefore, no pre-task planning time will be offered or dual tasks will be required or more elements will be included. This will enable the design of this study to control for any other possible interactions between IR demands and other resource-directing and/or resource-dispersing variables.

Some studies have adopted *simple-medium-complex* continuum of IR (e.g., Robinson, 2007) and others used *no-simple-complex* IR (e.g., Ishikawa, 2008). However, this study will employ a binary model of IR (less IR vs more IR) to avoid the confusion of the unsystematic employment and operationalisation of the three-level continuums of IR (e.g., no IR, +IR, ++IR). One can criticise that the design of the 'no IR' condition will not guarantee that participants will not reason during language performance. Moreover, due to the complexity and the abstractness of the construct of IR, there is no clear cut-off between +IR vs ++IR when it comes to the amount of IR required in each task as proposed by previous studies. Therefore, employing

a binary design, i.e., -IR vs +IR rather than an IR continuum will ensure that the variation between the two conditions in their IR requirements are salient and distinctive enough to result in variations in speech performance and perceptions of task difficulty, and hence learners' individual differences are more likely to play a more key role during performance.

What the literature review revealed is also supported by Jackson and Suethanapornkul's (2013) meta-analysis of TC. They highlighted two key issues that need to be tackled in studies examining the effects of reasoning on language production: 1) the scarcity of studies exploring the various elements of reasoning, and 2) the absence of reliable frameworks to conceptualise and operationalise reasoning. Furthermore, a research synthesis on TC by Malicka and Sasayama (2017) has revealed inconsistent findings with regards to previous research testing the predictions of the CH in terms of the effects of TC variables including reasoning on L2 performance. The inconsistent findings which Malicka and Sasayama have mentioned belong to studies investigating the same TC factor which is the case at the moment with studies exploring IR. Considering the inadequate definitions and the unsystematic investigation of IR in previous research, the current study attempts to handle these limitations and fill this research gap, while taking into account testing the predictions of the CH in terms of the effects of TC manipulated by IR on L2 learners' oral performance and perceptions of task difficulty.

2.9 Individual differences and language performance

Within SLA theories, the role of learners' individual differences in facilitating or hindering language learning and acquisition has been debated. A number of researchers are supportive to the standpoints that L2 acquisition is similar to L1 acquisition (e.g., Krashen, 1989). Thereby, this standpoint assumes that learning an L2 involves unconscious processing of TL input and is not highly dependent on any variation between learners in their individual differences. On the other hand, other researchers (e.g., Dörnyei, 2006; N. C. Ellis & Larsen-Freeman, 2006) believe that the processes involved in L1 and L2 acquisition are not identical and that learners' individual differences have different roles to play in developing both L1 and L2. Therefore, the tenets of this standpoint advocate that variation in learners' ability, cognitive and affective factors are assumed to contribute to the quality and the quantity of language performance and development.

Considering TBLT context, the CH argues strongly that L2 learners count more on their individual variables as complexity of language tasks increases (Robinson, 2001). Therefore, the

CH predicts that individual variables like working memory, language proficiency, aptitude, motivation and anxiety moderate the effects of TC on language performance and contribute to learners' perceptions of TD (Robinson, 2007). Furthermore, Robinson (2011a) debates that the cognitive ability variables (e.g., working memory, reasoning, aptitude) "as contributing to perceptions of TD would interact with characteristics of tasks contributing to their TC, inhibiting or promoting successful adaptation" (p.24) of these variables, and hence facilitate or hinder L2 performance and development. This suggests that learners' individual differences are associated with increased TC, and that the extent of their effects are more dependent on the level of TC. Consequently, it is inevitable for studies investigating TC to simultaneously explore its possible interaction with variation in individual factors. Responding to this research need, Study Two attempts to examine whether increasing TC along IR demands interacts with the variations in learners' language proficiency and working memory.

2.10 Language Proficiency

Language proficiency (LP) is defined as "Knowledge of language and the ability to access, retrieve and use that knowledge in listening, speaking, reading and writing" (Hulstijn, 2015, p. 21). According to Gaillard and Tremblay (2016), LP is "the linguistic knowledge and skills that underlie L2 learners' successful comprehension and production of the target language" (p.420). However, it can be argued that linguistic knowledge only does not guarantee a sufficient use of language, and thus non-linguistic competencies need to be considered as components of LP. Therefore, LP is likely to be associated with learners' L2 implicit knowledge and automaticity in using their linguistic knowledge in successful communication in the TL (Hulstijn, 2012).

From a research perspective, LP can be regarded as an individual difference that cannot be manipulated, and that any variations in learners' LP can be employed to describe any variations in the dependent variables under investigation (Hulstijn, 2015). Consequently, LP has been investigated widely as a factor that affects language learning and acquisition negatively or positively in a direct or indirect way. However, other studies which are not concerned with the effect of LP differences tend to control for its effect by identifying the levels of LP which enable researchers to choose participants with homogeneous LP through administering standardised or unstandardised instruments to measure LP. Thereby, measuring LP is a central issue for L2 pedagogy and research.

2.10.1 Measuring language proficiency

Viewing LP from a testing perspective, it is assessed in SLA research to group learners into homogeneous levels based on their global L2 proficiency or on a specific domain or skill (e.g., vocabulary size, pronunciation). However, due to the lack of consensus on systematic integrative measurements of LP, a variety of methods and tools have been developed to respond to the shifting requirements of SLA research and pedagogy. Historically, methods of LP assessment were first driven by a structural approach to language testing which involved testing discrete language structures (e.g., grammar, vocabulary) in isolation (Carroll, 1968). This was followed by advocating for an integrative approach to language testing (Oller, 1976) which considered language as including only integrated capacities rather than isolated elements. Following the limitations of these two approaches and the expansion of communicative and cognitive approaches to SLA, LP assessment was developed to measure “the comprehension and production abilities that L2 learners develop across linguistic domains (e.g., lexical competence, grammatical competence, discourse competence) and modalities (spoken and written) to communicate” (Tremblay, 2011, p. 340) more effectively in real-world situations. Hence, new standardised tests were designed to describe both oral and written competences and performances of the test-takers in the TL.

Current studies measuring LP adopt the scores of standardised tests that are recognised globally (e.g., IELTS and TOEFL) or administer original or simplified versions of standardised placement tests (e.g., Oxford Placement Test, Cambridge University Test, Michigan University Test). Most of these tests allocate learners to levels that correspond to either the Common European Framework of Reference for Languages (CEFR) or the American Council for the Teaching of Foreign Languages (ACTFL). Assessment of LP also involves using other tests which are assumed to tap into global proficiency (e.g., cloze tests and vocabulary size tests). Other instruments target oral proficiency through standardised or unstandardized speaking tests or oral interview tests. Elicited imitation tasks have also been adopted widely in L2 research as a measure of global oral proficiency (R. Ellis, 2009b; Erlam, 2006). However, other L2 studies only report the participants’ school or university internal scores as a reference to their LP. Other studies consider length of TL learning, exposure to the TL, or residence in the TL context as a tool to estimate learners’ levels of LP. As will be explained in more detail later, Study One considers the participants’ internal school assessment of their levels of LP, whereas Study Two

employs Oxford Placement Test and an elicited imitation task to group the participants into levels based on the CEFR.

2.10.2 Language proficiency and language performance

Considering a psycholinguistic viewpoint, LP is assumed to play a major role in processing L2 performance since L2 speech production is dependent on conscious and controlled attention (Kormos, 2011) , and hence it can drive or hinder language development. The amount of automaticity in L2 encoding and ability to allocate attention during performance appear to be associated with L2 speakers' LP. Research on the interaction between TC and LP attempts to explain whether variations in LP leads to variations in allocating attention, controlling existing interlanguage, and monitoring speech production. By maintaining a smooth flow of linguistic resources needed during performance, higher levels of LP are argued to back L2 processing by assisting learners attain parallel processing, and thus free more attentional resources to attend different aspects of performance (Kormos, 2011). It is hence predicted that high-proficiency learners are likely to operate better during heavy processing and storage access in more cognitively demanding L2 tasks, because the linguistic requirements which involve both storage and processing, are easier for them. In other words, it takes longer for high-proficiency speakers before the trade-off breaks down which is not available for low-proficiency speakers who, as cognitive load increases, are expected to suffer trade-off breakdown earlier, since they are processing both cognitive and linguistic demands (Wright, 2010). Thus, high-proficiency learners with high working memory will be privileged to produce better L2 performances on more complex tasks. This can lead to predict a possible interaction between language proficiency, working memory and task complexity. However, no study has investigated the interaction between LP, WM and IR as a TC variable in a TBLT context, which is a gap, this study strives to fill.

2.10.3 Language proficiency and Task Complexity research

Given TC research, there has been a tendency to control for LP as an individual factor rather than examine its interaction with the different variables of TC. Only few studies considered investigating the main effect of LP on language performance or development and its interaction as an independent variable with manipulating TC (Declerck & Kormos, 2012; Gilabert, 2007a; Gilabert & Muñoz, 2010; Ishikawa, 2006; Kuiken & Vedder, 2008; Malicka & Levkina, 2012). Gilabert (2007a) investigated the interaction between TC and LP on self-repairs. Gilabert

manipulated TC along *Here and Now* (narrative task), *number of elements* (giving-instruction task) and *reasoning* (decision-making). LP levels of the 42 participants (low and high) were measured by means of X-lex and Y-lex vocabulary size tests. Gilabert found no relationship between the learners' self-repair behaviours across the complex tasks and their LP. As LP is the focus of this section as a main independent variable, it is worth mentioning that this study used only one measure to test LP and did not include another standardized test of LP (e.g., Oxford Placement Test) for more reliable measurements of global proficiency.

However, Malicka and Levkina (2012) employed both X-lex and Y-lex and Oxford Placement Test to measure their 37 participants' LP to investigate whether LP regulated the effects of TC. Malicka and Levkina examined two TC variables, i.e., *-/+reasoning* and *+/-elements* through instruction-giving tasks. The findings revealed that the high-proficiency group produced more syntactically and lexically complex and accurate performance on the complex tasks, whereas the low-proficiency group produced more fluent language, while complexity and accuracy were unaffected. While the results for high-proficiency group fully supported the CH, the performance of the low-proficiency participants indicated that they might have prioritised attending fluency over complexity and accuracy suggesting a trade-off effect. This would mean that LP levels (low and high) interacted differently with TC requirements according to Malicka and Levkina.

Declerck and Kormos (2012) studied whether variation in LP would moderate the effects of increasing TC along dual task demands on L2 encoding and self-monitoring by using network descriptive tasks. Oxford Placement Test was employed to divide the 20 participants into two LP groups, i.e. B1 and C1. The findings indicated that the dual task demands influenced speech performance negatively but differently across the LP groups. The findings revealed that the C1 group outperformed B1 in terms of accuracy, speech rate and repairs, whereas no difference was found with regards to filled pauses. Declerck and Kormos concluded that repairs behaviour was not affected by variation in LP. Furthermore, they suggested that speech rate as a measure of fluency is more sensitive to differences in LP than pausing behaviour. The latter was assumed to be more associated with L2 learners' individual strategies which they employ while speaking regardless of their LP.

Ishikawa (2006) investigated whether LP interacted with the effects of TC manipulated along *+/-Here and Now* on L2 learners' narrative writing. The 52 participants were divided into two LP groups, i.e. low and high by employing Michigan English Placement Test. The results

detected main effects for LP on all language performance aspects, whereas TC was found to influence all aspects except lexis. However, the interaction results revealed that TC and LP affected structural complexity and fluency independently, but there were signs of interaction regarding accuracy (target-like use of articles) and lexis (TTR). Kuiken and Vedder (2008) also examined the interaction between the effects of TC and variation in LP on L2 learners' writing performance. TC was manipulated by increasing the number of elements, and LP was assessed by using a cloze test. The participants who were Dutch learning Italian (N = 91) and French (N = 76) as foreign languages wrote advice letters about holiday destinations. Though, the findings revealed main effects of LP on grammatical complexity, accuracy and lexis, no interaction effects were found between TC and LP on any aspect of the participants' writing performance.

2.11 Working Memory

Working memory (WM) is investigated within SLA research as one of the cognitive individual differences that can be associated directly or indirectly with language processing and performance (Skehan, 2015b). Recently, there has been a growing interest in TBLT context in exploring the roles WM play in L2 performance and development as a predictor or a mediator, and more specifically its interaction with TC (Baralt, 2015; Gilabert & Muñoz, 2010; Kormos & Sáfár, 2008; Kormos & Trebits, 2011; Mitchell, Jarvis, O'Malley, & Konstantinova, 2015; Mota, 2003). The devotion to incorporating WM in TC studies stems from the notion that WM is one of four key cognitive factors which are at stake during performing L2 complex tasks, i.e. declarative and procedural knowledge, speed processing, in addition to WM (Kyllonen & Christal, 1990). Furthermore, WM is argued to be at the heart of this line of research as a learner factor for its potential impact on regulating and managing L2 learners' linguistic repertoire and attentional resources during language performance (Wen, Mota, & McNeill, 2015).

WM is defined by Atkinson and Shiffrin (1968) as a provisional system for the long-term memory, in addition to its leading role during performing high cognitively complex tasks that require comprehending, reasoning or information manipulation. This definition considers WM as a gateway to long-term memory, and the key to this gate is the quantity and quality of repetitions or rehearsal of any information stored temporarily in the short-term memory. Adding a new dimension to WM, Engle, Tuholski, Laughlin, and Conway (1999) postulate that WM facilitates both storage and attention which are assumed to suffer limitations in their capabilities. With storage-attention limited capacity in mind, the role the central executive plays in

regulating both storage and attention is crucial in determining WM effectiveness during performing complex tasks. From an information-processing perspective, (Conway et al. (2005)) define WM as “a multi-component system which is responsible for active maintenance of information in face of ongoing processes and/or distraction” (p.770). This definition backs the fact that WM acts as a workstation for processing during demanding tasks which causes an intensive flow of information that needs more processing resources resulting in breakdown or trade-off in language performance and/or distraction of attention. It is hence anticipated that these processing challenges cannot be resolved without the assistance of the WM system.

Building on the aforementioned definitions, Baddeley (2003) defines WM as “a limited capacity system, which temporarily maintains and stores information and supports human thought processes by providing an interface between perception, long-term memory and action” (p.829) and that it is “assumed to be necessary for a wide range of complex cognitive activities (Baddeley, 2012, p.7). Baddeley’s definition brings attention to the main features, components and functions of WM. This definition suggests that WM has different components that are limited in capacity but in collaboration while operating. According to Baddeley (2015), this system functions in coordination with the long-term memory to support both cognition and language. Baddeley further advocates that WM has prime implications for language processing and that any deficiency in WM is assumed to affect language processing, performance and development negatively. However, some researchers claim that WM does not form a multicomponent system and that it is considered part of a single entity called ‘long-term WM’ (e.g., Ericsson & Kintsch, 1995). Hence, WM can be seen as indivisible system that feeds in human intelligence system (Cowan et al., 2005). Despite the standpoints that consider WM as a unitary model, the multicomponent model of WM proposed by Baddeley and Hitch (1974) has been adopted widely within SLA and TBLT research.

2.11.1 Baddeley’s model of working memory

Supported theoretically and empirically, Baddeley and Hitches (1974) put forward a three-component model to help researchers understand how the human memory system functions during cognitive activities including language tasks. It is thereby hoped that the availability of such a framework will advance research on L2 learning and acquisition with potential pedagogical implications in the fields of SLA and TBLT. The three components of Baddeley and Hitchens’s WM model are: 1) *the central executive*, 2) *the phonological loop*, and 3) *the visuospatial sketchpad*. A fourth component has been included later to the model, i.e. *the*

episodic buffer based on empirical evidence from language impairment research (Baddeley, 2000). The four components of Baddeley’s WM model as shown in Figure 4 below will be explained in more detail.

The *central executive* is considered the heart of WM model, the most complicated component, and the least understood (Baddeley, 2003). Beforehand, it was assumed to jointly regulate storing information and processing attention. Empirical evidence has appeared later suggesting that storage is not a property of the central executive and that another component, i.e. *the episodic buffer* is supposed to be in charge of the storage role (Baddeley, 2015). Thus, the main role of the *central executive* is to act as a general processor of information by managing and directing the available attentional resources to attend or ignore certain aspects of the input (Baddeley, 2003). It is also suggested that the *central executive* assists in maintaining bidirectional flow of information to the long-term memory and the other slave systems (ibid). Though, it has a limited capacity, the central executive enables humans to consciously control and direct their attention to desired stimuli or away from undesired stimuli or distractions (Baddeley, 2000).

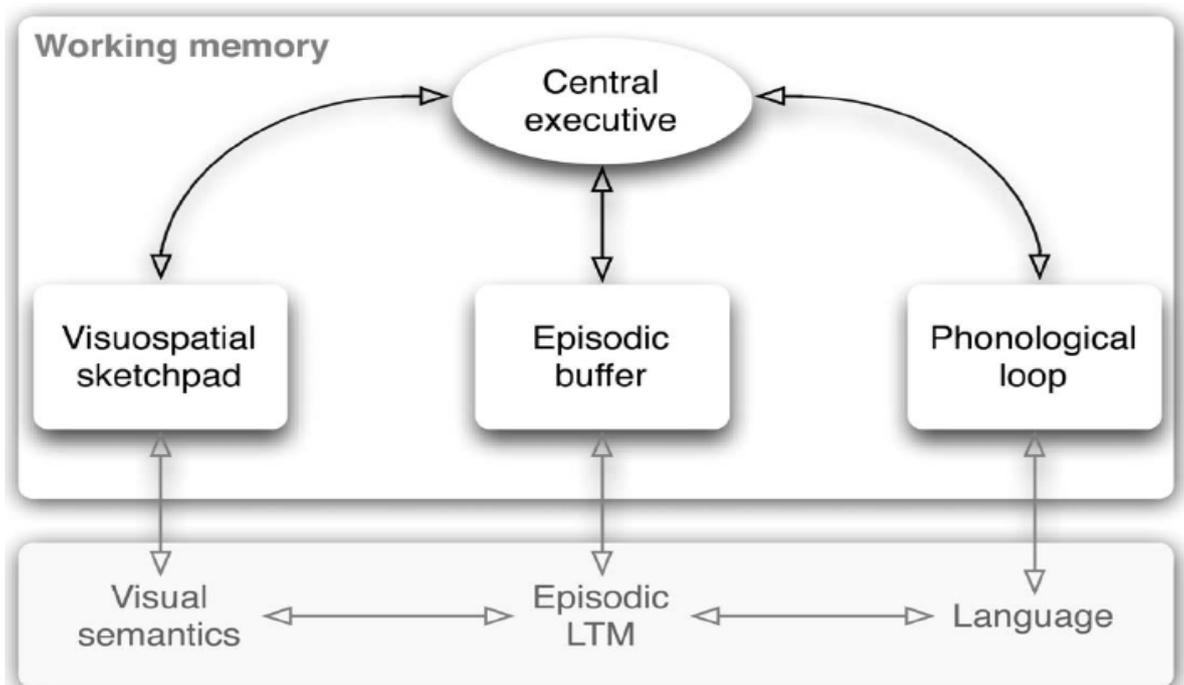


Figure 4. The Multicomponent model of WM (Baddeley, 2003, p. 835)

The *phonological loop* is the articulatory component of the WM model which deals with the verbal data. It comprises two subsections, i.e. the ‘*short-term phonological store*’ which keeps

phonemes for a brief time, and the ‘*articulatory rehearsal*’ which facilitates rehearsing the stored phonemes (Baddeley, 2003). The phonological loop can regulate both auditory-visual data, encode them as verbal information, and then store them temporarily in the phonological store. In case that the rehearsal component provides sufficient opportunities for rehearsing, repeating or using these phonemes, learning or acquisition can take place. Accordingly, these new phonemes will be transferred to the long-term phonological memory. This is why, the phonological loop is expected to be a reliable indicator of learners’ aptitude to learn or acquire a second language successfully (Baddeley, 2012). According to Baddeley (2015), there is “a clear link between the phonological loop and phonological long-term memory, a link that operates in both directions” (p.19) which evidences that the phonological loop can aid learners to acquire new linguistic items faster and more effectively than others. It is no wonder therefore that the phonological loop has attracted more attention than any other WM components in SLA research. It worth mentioning that the phonological loop can be assessed by using WM tests that require instant ongoing recall of items, i.e. digit or words.

The *visuospatial sketchpad* is the component that is responsible for receiving and storing visual and spatial data. In other words, it can store and manipulate visual information about the characteristics of objects, their movements and locations within the surrounding context (Baddeley, 2000). This suggests that the visuospatial sketchpad enfolds two separate subcomponents: one for dealing with visual information, i.e. *the visual cache* and another for processing spatial information, i.e. *the inner scribe* (Deyzac, Logie, & Denis, 2006). Regarding SLA, the visuospatial sketchpad is supposed to contribute in learning and acquiring the semantic features of new words by linking their meanings with specific visual representations of these words (Baddeley, 2003). However, this component has received little attention in SLA research due to the fact that its direct impact on L2 performance and development has not been confirmed yet, compared to the contributions of other components, i.e. *phonological WM* and *executive WM* (Wen, 2015).

The *episodic buffer* is the fourth and lately included component to the WM model. It is managed by the central executive and acts as a hub that incorporates all information coming from the other two slave systems, i.e. the phonological loop and the visuospatial sketchpad. In other words, the *episodic buffer* is considered as “a multidimensional storage system, capable of combining information from the visuospatial and verbal subsystems and linking it with further information from perception and long-term memory” (Baddeley, 2015, p.20). The episodic

buffer then integrates all kinds of information into unified episodes that can be accessed and processed consciously (ibid).

2.11.2 Measuring working memory

The issue of measuring WM is still open to debate among researchers and practitioners within cognitive psychology and SLA due to a lack of agreement on the reliability and validity of the current tools that measure the capacity of learners' WM (Mitchell et al., 2015). However, researchers tend to choose a WM task or a battery of tasks that serves the aims of their research and taps into the WM component(s) under investigation. Several WM span tests have been designed to serve a purpose of assessing the test-takers regarding their simultaneous capacity of information processing and memory storage (Daneman & Carpenter, 1980).

Conway et al. (2005) advocate that “WM span tasks, such as the counting span, operation span, and reading span tasks, are among the most widely used measurement tools in cognitive psychology” (p.769). These tasks are also popular in SLA studies which are interested in exploring the moderation effects of WM on L2 performance and development. Other WM span tasks that are suggested by cognitive psychology include the speaking span task (Finardi & Weissheimer, 2009) and the listening span task (Goo, 2010). The verbal WM span tests are administered individually in counterbalanced protocols using randomised sets of items in order to minimise any possibility to assist the test-takers to adopt certain strategies or patterns that can affect the score of their spans. All these tasks contain processing components (require memory processing) followed by items to be recalled (require memory storing) and at the end of each set, the ‘*to-be-recalled items*’ are articulated spoken or written. The tests are built of sets with increased number of items to represent finally the participants' WM spans.

The Reading Span Test (Daneman & Carpenter, 1980) requires the task-taker to read and judge loudly the appropriateness of a set of sentences and then recall the final word of each sentence in order. The test begins with a two-sentence set and ends at a six-sentence set. The participant's WM span is determined based on the last set in which he/she successfully recalls the last words of all sentences of that set. An automated version of the reading span test has been produced by (Unsworth, Redick, Heitz, Broadway, & Engle, 2009).

The Operation Span Test (Turner & Engle, 1989) includes repetitive sets of processing items (judge correctness of answers of arithmetic problems) followed by storage items (numbers,

letters or words to be recalled). The operation components are designed to inhibit the participants from rehearsing the storage items. The items increase gradually in each following set, and the participants are finally required to recall only the items that are given after each mathematical problem.

The Counting Span Test (Case, Kurland, & Goldberg, 1982) is a computer-based task which requires the participants to count shapes of certain colours which are displayed simultaneously with other shapes of similar colours. The colourful backgrounds and other similar shapes serve as visual distractors for the participants who are finally requested to say loudly the count of shapes for each set in the same order they were displayed. Other verbal WM tests include digit span tests which are of two types, i.e. *Forward-Digit Span* (Botwinick & Storandt, 1974) which measures storage only and *Backward-Digit Span* (Kormos & Sáfár, 2008) which measures both storage and processing. The latter is adopted in this study, and therefore will be explained in more detail in the Methodology Chapter of Study Two.

2.11.3 Working memory and language performance

According to Baddeley (2015), it is assumed that certain components of the WM model are linked to learning and acquiring first and second languages. Hence, Wright (2015) suggests that WM has recognised implications for L2 performance and development since WM “supports complex management of verbal input and retrieval and will therefore aid sentence processing, reading comprehension and general language fluency” (p. 290). Given language processing, Skehan (2015b) defines WM as a buffer for language input and output, and “a workplace for the solution of problems which occur during input processing, output processing or general analysis or planning” (p.189). According to (R. Ellis, 2015), WM is “a mental construct that accounts for how the key processes of perception, attention and rehearsal take place. It is believed to play a central role in L2 acquisition” (p. 983). Taking also another language processing perspective, Wen (2012) identifies WM as “the limited capacity of multiple mechanisms and processes in the service of complex L2 activities or tasks” (p.10).

The abovementioned definitions accentuate the significant roles WM can play not only in maintaining successful L2 processing and performance, but also in effectively learning and successfully acquiring new language items. The definitions confirm the limitation of WM capacity and its mediation of TC. However, L2 speech production is argued to be effortful and serial due to the less rich, less organised, and less accessible mental lexicon (Kormos, 2006) ,

and hence WM is anticipated to be heavily taxed by the pressing needs to allocate attentional and memory resources to specific stages of speech production (Skehan, 2015b). During processing L2 performance, WM is supposed to aid and regulate language requirements at both the Conceptualisation and Formulation stages of speech production (Finardi & Weissheimer, 2009) by directing and dividing speakers' attentional resources through its central executive to attend form and/or meaning (Baddeley, 2012).

The Limited Attentional Capacity argues that attention is not expandable, and therefore more demanding tasks are expected to result in consuming memory and attentional resources (Skehan, 2014), with negative consequences on storage capacity and speed processing of WM (Mackey, Adams, Stafford, & Winke, 2010). On the other hand, the Cognition Hypothesis claims that even though WM has a limited size, attention is expandable to attend lexical, morphological and syntactic aspects of L2 system, and monitor and revise the required semantic-pragmatic concepts (Robinson, 2011a). As a result, the expansion of attention can compensate for WM limitations, and therefore the negative effects of TC on language performance will be only restricted to fluency. This means that WM can still aid learners in directing attention, accessing the mental lexicon, and retrieving required syntactic and lexical language (Kormos, 2011). This leads to anticipate that WM as a cognitive ability factor is more likely to be at stake as tasks increase in their cognitive demands.

2.11.4 Working memory and TBLT research

Previous research in general has investigated WM as an individual factor responsible for regulating attention (Gathercole, 1999), processing information (Baddeley, 1986), performing cognitive activities (Baddeley, 2003), and reasoning (Salthouse, 1992). Researching WM as a variable that mediates L2 learning and development or interacts with other factors affecting this process has recently become a trendy line of research to the extent that special volumes are dedicated only for WM-SLA joint research (e.g. Wen, Mota, & McNeill, 2015). With reference to TBLT research, some studies investigated the relationship between WM and L2 development (e.g., Kormos & Sáfár, 2008; Mitchell et al., 2015), whereas other studies explored the role of WM in L2 performance (e.g., Ahmadian, 2013; Gilabert and Munoz, 2010; Kormos & Trebits, 2011; Mota, 2003).

In a correlational study, Gilabert and Munoz (2010) explored whether variation in WM would explain variation in LP and L2 performance. A reading span test was administered to measure

the participants' WM, while three tests were used to assess LP, i.e. Oxford Placement Test, two vocabulary size tasks, and a phonetic classification task. The 59 participants were grouped into low-high groups based on their LP and they performed a video-based narrative task. Speech performance was measured by S-nodes per AS-unit (syntactic complexity), Guiraud's Index (lexis), errors per 100 words (accuracy) and unpruned speech rate (fluency). Though Gilabert and Munoz did not find any correlation between WM and LP, they found that WM correlated with lexical complexity and fluency. Moreover, LP was found to correlate with all aspects of oral performance except syntactic complexity. In terms of LP groups, only a medium correlation was detected between WM and lexis for the high LP participants. These results suggested that LP as a predictor of lexical complexity was more reliable than WM. Gilabert and Munoz predicted that WM would influence speech performance on tasks with increased TC. However, since the data in this study were collected from performing only one task, it was not feasible to test these predictions. Another limitation of this study was using only one measure to operationalise each dimension of speech performance which might have missed depicting a more comprehensive account of the findings.

Mitchell et al. (2015) were also interested in the relationship between WM and LP on their effects on L2 processing and development. Based on their TOFEL scores, the 36 Chinese learners of English were grouped into three levels of proficiency (beginner, intermediate, advanced). Their WM was measured by administering an operation span task and forward-digit span tasks in L1 and L2. The participants performed elicited imitation and reading tasks in English to measure their language processing and production. The findings did not show any correlation between LP scores and L1 digit span and operation span scores, whereas LP correlated with the L2 digit span scores. The study found stronger relationship between WM and LP in the case of high proficiency levels. The results indicated that LP interacted differently with L1 and L2 digit span scores suggesting that these tasks are not language-independent. This conclusion was in contrast with previous findings which found correlational patterns between L1 and L2 digit span tasks (e.g., Wright, 2010). However, it could be argued that such a result cannot be generalised to languages that are different from Chinese (e.g. Arabic). Another interesting conclusion of this study was finding that the participants of different LP levels were utilising their WM system differently suggesting that the phonological component was more beneficial for low-proficiency learners, whereas the high-proficiency learners were more dependent on their WM central executive.

Another study examined WM and L2 development was Kormos and Sáfár (2008) which focused on the effect of one component of WM, i.e. phonological loop on LP and L2 development in terms of the four language skills and use of English. WM was measured by a backward-digit span and non-word tasks. The 121 participants were divided into two LP groups (beginners and intermediate) by administering Cambridge First Certificate Exam. The findings confirmed a correlation between the phonological memory spans (non-word test) and LP (Cambridge test) and L2 development (writing and use of English) in case of the intermediate group only. Another correlation was spotted between WM scores (backward-digit task) and LP (Cambridge test) and L2 development (all language skills). The findings suggested an interaction between the phonological short-term memory and LP. These results led to conclude that the phonological loop and the central executive were discrete components of the WM system which was confirmed through the non-significant correlation between the scores of the phonological memory test as a measure of the phonological loop and the backward-digit span test as a measure of the central executive.

Kormos and Trebits (2011) investigated the relationship between variation in WM and increased TC on L2 oral performance. A backward-digit span task was employed as a measure of WM. The 44 participants performed two narrative tasks with increased TC, i.e. tell a story (simple) and invent a story (complex). To operationalise syntactic complexity, different measures were considered to tackle subordination, length of clause, and use of different verb forms. A measure of diversity, i.e. D was adopted to measure lexical complexity, ratio of error-free clauses to tap into accuracy, and speech rate as a measure of fluency. It was found that high WM benefited syntactic complexity in terms of ratio of subordination and mean length of clause. However, the effect of WM was only observed in the simple task. In terms of the effects of TC, it was revealed that the complex task produced more accuracy but less lexical complexity, whereas no effect was detected on grammatical complexity of fluency. This study did not employ a standardised test to measure the participants' LP and adopted the teachers' judgment which might affect the results.

In a small-scale study (N = 13), Mota (2003) investigated the relationship between WM and L2 speech performance as operationalised by a number of CALF measures. The participants' WM was assessed by using a speaking span task, whereas the researcher adopted the participants' self-reporting of length of studying English as a measure of LP. They performed a descriptive task based on a picture strip and a narrative task which required them to retell the story of a

movie they liked. The two oral tasks were monologic but differed in the time allocated to perform each of them, i.e. descriptive task (two minutes) and narrative task (open time). Syntactic complexity was measured by number of dependent clauses, accuracy by errors per 100 words, lexical complexity by weighted lexical density, and fluency by speech rate, number of pauses and hesitations, and mean length of run. Mota hypothesised that WM would correlate positively with fluency only in terms of speech rate and length of run. Also, positive correlations were predicted with respect to syntactic and lexical complexity, whereas accuracy was expected to correlate negatively with the scores of the WM test. In line of the predictions, WM scores correlated positively with fluency and syntactic complexity and negatively with accuracy. Contradictory to Mota's predictions, WM correlated negatively with lexical complexity. This study advocated WM as measured by the speaking span test as a reliable predictor of syntactic complexity, accuracy and fluency but not lexical complexity. This study was questioned for its small number of participants ($N = 13$) and for not controlling for LP as a possible mediator of L2 performance. Furthermore, using two tasks that were assumed to pose equivalent cognitive demands and not restricting the time allocated to the narrative task could have misled the reported findings.

2.12 Conclusion

This chapter presented a review of task and TBLT, followed by a discussion of the models of speech production and aspects of L2 performance. The chapter also presented a thorough investigation of task complexity and its theoretical frameworks. A full discussion of intentional reasoning was also presented, including definitions, frameworks and studies of IR. The final sections of this chapter were dedicated to discussing the issue of individual differences in task-based research with a focus on language proficiency and working memory which will be included to the investigation of Study Two.

The literature reviewed in this chapter revealed different theoretical and methodological issues that were worth highlighting. The absence of specific criteria for the choice of CALF measures in studies that investigate TC is an issue that needs further investigation. This is particularly important as a less than systematically monitored choice of measures may result in misleading findings. Drawing on different theoretical frameworks, researching TC could be regarded as the area that still needs more attention and collaborative efforts. The diverse assumptions of the two largely used theoretical models of TC, i.e. Cognition Hypothesis (Robinson, 2007) and

Limited Attentional Capacity (Skehan, 2009) have invited plenty of research often leading to inconsistent conceptualisation and operationalisation of TC variables, and occasionally resulting in findings that fail to lend full support to any of the two models.

The different ways each model perceives the relationship between task complexity and task difficulty adds more ambiguity to how these two constructs need to be addressed in task-based research. Moreover, the lack of a thorough investigation of TD that could help to identify factors that contribute to learners' perception of difficulty is one of the research gaps this study attempts to help fill. Therefore, an in-depth exploration of this construct can offer a valuable contribution to the development of an index of TD with significant implication for L2 pedagogy. However, another gap that is still present in previous research is the focus on examining TD as a dependent variable rather than an independent variable. While the need to explore how L2 speech performance is influenced by TD is a valuable research focus, for reasons of scope the current study will not investigate it.

Reviewing the literature further revealed that the CH offers a more feasible framework for operationalising the constructs of TC and TD than the LAC. Therefore, this study adopts the CH as the theoretical framework and attempts to test its predictions. The results of the study, although mainly analysed to test the assumptions of CH, can also be examined to see if they support principles of LAC. Reviewing previous studies has showed lack of empirical findings that fully support the CH claims. Furthermore, the vagueness of the definitions of TC factors including intentional reasoning is another gap this study aims to address. Adding to this is the paucity of studies that investigate IR and the absence of systematic frameworks that operationalise this construct.

Though, IR has been investigated sparingly in TC research, its interaction with learners' individual differences in language proficiency (LP) and working memory (WM) has not been researched before. The question of how LP and WM interact with IR effects on speech performance has not been investigated. This will be a prime aim of the current study. In an attempt to address the aforementioned issues and in order to fill the research gaps identified in the literature review, this study is designed carefully to 1) re-define and re-operationalise IR more systematically; 2) identify the factors that contribute to perceptions of TD; 3) examine the interaction between LP, WM and IR; 4) test the predictions of the CH; and 5) interpret the findings with caution by considering the effect of the choice of CALF measurement.

In the next chapters, the methodology, results and discussion of Study One will be presented. This will be followed by presenting the methodology, results and discussion chapters of Study Two. Finally, a conclusion chapter will draw final remarks from the findings of Study One and Two, highlight their contributions and implications, acknowledge their limitations, and suggest areas and issues to be addressed by future research.

Chapter 3: METHODOLOGY: STUDY ONE

3.1 Introduction

This chapter presents the aims of Study One, the research questions, the study design, and the methodological procedure. First, the aims and the research questions of this study will be presented. Next, the study design employed here will be explained and the choice will be justified. Then, a description of the participants and their recruitment is presented, followed by the tasks, instruments and the pilot study. The ethical procedure is then reported, followed by a detailed description of data collection procedure. The choice of the measures of analysis is acknowledged and then justified. The procedures of transcribing and coding the data will be discussed, and finally, a summary is made of the statistical analyses used.

3.2 Aims of the study

Study One aimed to investigate the effects of manipulating Task Complexity (TC) on L2 learners' oral performance and perceptions of Task Difficulty (TD). For TC in this study, a novel perspective was taken of manipulating two levels of intentional reasoning, i.e. -IR demands versus +IR demands. Language oral performance was operationalised through measures of syntactic complexity, accuracy, lexical complexity, and fluency (CALF). In addition to these linguistic analyses of complexity, accuracy, fluency, the study also included a novel dimension of learner perception to fully contextualise learner performance analyses. The participants' perception of TD was measured by administering a retrospective questionnaire. The present study thus aimed in an original way to test hypotheses arising from the claims of the Cognition Hypothesis (Robinson 2001, 2005, 2007, 2011) about the positive effects of increasing TC on complexity and accuracy, and the negative effects on fluency.

3.3 Research Questions (RQ) & Hypotheses (H)

RQ1: Does manipulating TC through the amount of IR required affect learners' L2 oral performance?

This question was divided into four sub questions to respond to each aspects of the participants' oral performance, i.e. syntactic complexity, lexical complexity, accuracy and fluency.

RQ.1a: Do (-/+) IR requirements affect the syntactic complexity of oral performance?

H.1a: Performing monologic narrative tasks with +IR will be associated with more syntactically complex language than performing tasks with -IR.

RQ.1b: Do (-/+) IR requirements affect the lexical complexity of oral performance?

H.1b: Performing monologic narrative tasks with +IR will be associated with more lexically complex language than performing tasks with -IR.

RQ.1c: Do (-/+) IR requirements affect the accuracy of oral performance?

H.1c: Performing monologic narrative tasks with +IR will be associated with more accurate language than performing tasks with -IR.

RQ.1d: Do (-/+) IR requirements affect the fluency of oral performance?

H.1d: Performing monologic narrative tasks with +IR will be associated with less fluent language than performing tasks with -IR.

The hypotheses are set following the predictions of the Cognition Hypothesis (Robinson 2001, 2005, 2007, 2011), that manipulating TC along resource-directing dimensions, i.e. (-/+) IR, is assumed to direct L2 speakers to focus on form; the task should thus promote complexity and accuracy when the task is more complex (+ IR), while a focus on meaning is expected when the task is simple (-IR), thus promoting fluency.

RQ.2: Do L2 learners perceive the more complex task that requires IR as more difficult?

H.2: L2 learners will perceive the more complex task (+IR) as more difficult than the less complex task (-IR). This hypothesis is formed to test standard assumptions about the demands felt by learners about the levels of task complexity vs task difficulty (Robinson, 2001; Tavakoli, 2009a; Tavakoli & Skehan, 2005) . This question is addressed by a retrospective questionnaire on how the participants would perceive the two tasks in terms of their level of difficulty. This question is further addressed by collecting qualitative data about why the participants have perceived a certain task as more difficult than the other. Hence, the qualitative data collected from the task difficulty questionnaire are carefully analysed and interpreted to identify what factors contribute to the perceptions TD.

3.4 Study Design

As noted, Study One was designed to investigate the effect of TC manipulated by two levels of IR on L2 oral performance and perceptions of TD. In order to develop a more in-depth understanding of the relationship between the variables under investigation, answer the research questions from different perspectives, and provide research validity (Dörnyei, 2007), a mixed-method, within-participants study design was used. The design was counterbalanced between participants to minimise the effects of order and practice on task performance. Each participant performed two tasks, each with a different degree of IR, and performance in one task was compared against performance in the other task.

The mixed-method design enabled this study to integrate quantitative and qualitative data for triangulation purposes (Creswell, 2015). By means of combining quantitative and qualitative tools of data gathering, this study would achieve higher levels of validity as each model can help enhance the strengths of the other model as well as overcome any impending shortcomings in each one (Dörnyei, 2007). Furthermore, this design would empower the current study to gain a more in-depth understanding of the constructs under investigation by utilising a multi-level exploration of the collected data (*ibid*). Since this small-scope study comprised only two conditions and one group of participants, employing a within-participants (repeated-measures) design was the ideal option since the same measures were taken for each participant on two different occasions. Dealing with one group has helped this study avoid the complications of other designs (e.g., between-participants), which requires more participants with homogenous characteristics (Rasinger, 2013), a requirement which is arguably more difficult to control (Litosseliti, 2010).

By employing a within-participants design, the current study needed fewer participants to achieve the desired outcome, which would not be the case if the design was between-participants (Creswell, 2015). Hence, recruiting a smaller sample size in this study, reduced time and efforts and increased the possibility to control for the individual differences between the participants. According to Sauro and Lewis (2016), within-subjects designs require smaller sample size compared to between-subject designs, i.e. 1:3 to attain the same desired statistical confidence level. In other words, recruiting 20 learners as in this study using a within-subjects design will equal recruiting 60 participants in a study utilising a between-subjects design. This made the within-participants design an effective and practicable choice for this study.

The research design of the current study employed one independent variable with two levels, i.e. -IR and +IR. Syntactic complexity, lexical complexity, accuracy, fluency of the learners' oral performance served as the dependent variables and were operationalized through a variety of general measures. Including measures that represent the four aspects of oral performance guaranteed that the study would capture a more holistic picture of the impact of increased TC on the learners' speech production. The fifth dependent variable was the participants' perceptions of TD. Table 2 below shows the study design and the variables under investigation in Study One.

Table 2. The study design and variables of Study One

| Study Design | Independent Variable | Dependent Variables |
|---|---|--|
| <p>Mixed-method Within-participants</p> <p style="text-align: center;"></p> <p style="text-align: center;">One group N = 20</p> | <p>Task Complexity</p> <p style="text-align: center;"></p> <p style="text-align: center;">- Intentional reasoning + Intentional reasoning</p> | <p>Oral language performance</p> <p style="text-align: center;"></p> <p style="text-align: center;">Syntactic complexity Lexical complexity Accuracy Fluency TD perceptions</p> |

3.5 Participants

Twenty Jordanian secondary school students participated in this study. All the participants were male, aged 16, and with Arabic as the first language. This study recruited younger participants to respond to the need to target this age group in TC research which was overlooked by only focussing on adult learners at university level. All participants had studied English for about ten years at school and had never lived in an English-speaking country for more than six months before. Students at this school receive five sessions of English language per week (typically 4 hours). The school is one of the leading private schools in Jordan which offers extracurricular activities in English, and uses worldwide recognised language textbooks that follow the CEFR (Council of Europe, 2001). The students were assigned to different level classes that matched their English language proficiency. The placement of students into levels was based on their results of internal English language exams and continuous assessments that covered all

language skills. There were three levels of English language classes in this school, i.e. A, B and C. The participants in this study were randomly selected from the highest level (A), which is equivalent to B2 level of CEFR. At the time of data collection, the mean of their scores in English for the first semester was 91.2% and the standard deviation was 3.79 (see Table 3 below).

Table 3. Participants' scores in internal English exams and continuous assessments

| | N | Min. | Max. | Mean | Std. Dev. |
|--|----------|-------------|-------------|-------------|------------------|
| Participants' scores in English based on internal tests and continuous assessments | 20 | 85.00 | 99.00 | 91.20 | 3.79 |

Based on their overall scores, the participants were expected to be homogenous in terms of language proficiency level. As noted, this was equivalent to the CEFR B2 level. This suggested that they were independent users of English language, and thus they “can interact with a degree of fluency and spontaneity and can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue.” (CEFR, Council of Europe, 2001, p. 24). This descriptor is assumed to be ideal for the purpose of this study by ensuring that the participants have the ability to perform the two oral narrative tasks, cope with their increased cognitive and linguistics demands, and do not find the tasks overwhelming.

3.6 Tasks and Materials

The present study employed video-based tasks as stimuli for eliciting monologic oral narrative performance. Oral narrative tasks have been frequently used in CALF and TBLT research (e.g., Bygate, 2001; Ishikawa, 2011; Pang & Skehan, 2014; Tavakoli & Foster, 2008). They are prevalent tasks for both pedagogic and research reasons. Pedagogically, oral narratives are ecologically valid and form the cornerstone of L2 classroom interaction (Skehan, 2014). From a research perspective, oral narratives are reliable tasks since they allow researchers to collect data that facilitate a better understanding of the processes of SLA that are dependent on L2 production, by marking learners’ language ability developmentally and descriptively (Doughty & Long, 2000). Eliciting oral narratives based on wordless cartoons or films is assumed to “allow the researcher to constrain what the speaker will talk about while not putting specific words or sentence constructions into their mouth, and without making demands on memory”

(Segalowitz, 2010, p.43). By recruiting video-based tasks, the participants will find very limited opportunities to plan while speaking, which serves the purpose of the current study and controls for any possible effects of planning on language performance.

Following Skehan and Foster (1999) who used *Mr. Bean*, and Wang and Skehan (2014) who used *Shaun the Sheep*, this study employed *Pat & Mat* (Beneš & Jiránek, 1976), stop-motion animated series featuring two handymen who face self-made problems, trying to solve them in unpredictable and creative ways. The main two characters in this series, *Pat and Mat*, always get involved in problematic situations, but they never give up until they solve their problems by relying on their imagination and willpower. *Pat & Mat* has been selected for this study because a) it is based on silent videos and is an ideal stimulus to elicit speech production, b) the foreground and background of the scenes are rich with different events and, therefore, there are many actions to tell and describe, c) it offers fertile grounds for reasoning about intentions and thoughts to solve problems, and d) the challenging situations in the episodes have been sorted out in unpredictable ways and based on on-the-spot decisions.

The selection process of the video clips for this study considers De Jong and Vercellotti's (2016) framework of employing narrative tasks in L2 research that ensure using equivalent prompts to provoke comparable performances. De Jong and Vercellotti suggested that prompts should be chosen based on three similar elements, i.e. "narrative structure, storyline complexity, and number of elements" (p.402). Moreover, they argued that data about perceptions of task difficulty are valuable for validating the prompts selection. Therefore, judging which task is more complex should not be merely considered based on the outputs of the employed measures but also based on the participants' retrospective feedback (Levkina & Gilabert, 2012).

After careful piloting and examinations of a range of clips (as discussed in section 3.7), two clips were selected which had the same storyline complexity, duration, narrative structure, and number of characters. Furthermore, the choice was driven by the recommendation of De Jong and Vercellotti (2015) "to identify equivalent prompts that differ only in the feature under investigation" (p.388). Therefore, the two video tasks were only different at two levels, i.e. the amount of IR required and tasks instructions. This empowered the study to offer a more successful and systematic operationalisation of IR at the levels of content selections and task instructions. The selection was therefore in line with the study definition of IR as "reading other people's thoughts and understanding their desires, beliefs and motives to draw true conclusions about what they intend to do, why and what consequences follow." The participants were

required in the +IR clip to explain what the characters were doing or would be doing to solve their problematic situations and why by reading their minds and intentions, understanding their desires, and predicting the actions and their consequences.

Taking into account the aforementioned considerations, one video clip was assigned for the -IR condition, and another clip for the +IR condition. The instructions of the -IR task asked the participants only to *'tell and describe what is happening in the clip in details in English'*. By comparison, the instructions of the +IR task explicitly asked the participants not only to *'tell and describe what is happening'* but also to *'explain why the characters solve their problems or behave in certain ways, and explain their intentions and the unintentional consequences of their actions'*. The participants were therefore, fully encouraged to read the characters' thoughts, understand their beliefs and desires, explain their intentions, predict and justify their actions and explain their consequences. The more demanding requirements of the +IR conditions were assumed to push the participants to focus on form rather than meaning (as explained below) to express intentionality and draw true conclusions about the characters' actions and reactions. As a result, positive gains were predicted in the participants' language performance in terms of complexity and accuracy at the expense of fluency. The instructions were written both in English and in the participants' L1 (Arabic) and read aloud to participants to ensure full comprehension. In order to ensure that the narrative tasks appeared genuine, the two instruction rubrics incorporated a similar statement asking the participants to imagine that they are narrating the story of each clip to someone who is not watching the video. The rubrics of the tasks instructions are shown in Appendix 1 and 2.

The video of the -IR task showed Pat & Mat enjoying a sunny day when they decided to cook lunch outdoor, but suddenly it started raining. They kept trying to overcome this problem until they finally decided to cook inside their cottage, where they encountered a new set of challenges. The events of this video were clear and did not include unpredictable ideas or imaginative actions. Therefore, the clip did not require the participants to reason about the characters' intentions while narrating the story. Hence, the less demanding requirements of the -IR task would not direct the participants to attend form, but instead they would prioritise meaning, affecting the fluency of the performance positively.

The video of the +IR task was about Pat & Mat trying to fly their car like an airplane. They came up with weird but creative ideas to make their car fly. Participants were required to read Pat and Mat's thoughts and predict what they intended to do to make their car fly and why. The

story of the +IR condition was more imaginative and unpredictable which was assumed to demand more cognitive efforts in reading the characters' minds to reveal the uncertainty and draw conclusions based on the few available premises. Meeting these requirements during the narration while trying to keep up with the pace of the story, would impose extra linguistic and cognitive burdens on the performers. Those demands were expected to have a systematic impact on the complexity and accuracy of language performance, and also on perceptions of TD. Task materials included the ethics forms, i.e. an information sheet and a consent form (see Appendix 3), a language background questionnaire (see Appendix 4), and a retrospective questionnaire on the learners' perceptions of TD (see Appendix 5).

This retrospective questionnaire aimed to tap into how participants perceived the performed tasks in terms of difficulty. Following (Robinson, 2001; Tavakoli, 2009a), the TD questionnaire contained two scale-questions with four adjectives to describe each task as *very easy - easy - difficult - very difficult*, and two open-ended questions to justify the learners' judgement. Questionnaires are frequently used in applied linguistics research for their effectiveness in "gathering a large amount of information quickly in a form that is readily processible" (Dörnyei, 2007). In this study, the retrospective questionnaire as a data gathering tool was employed to stimulate the performers to immediately recall and verbalise their feelings and thoughts about the level of TD after performing the narrative tasks. The immediate stimulated recall of perceptions through questionnaires or interviews is assumed to be a reliable and valid technique to collect retrospective data as the relevant information is retrieved from the participants' memory on short time interval to minimise the effect of memory decline (Dörnyei, 2007).

Studying learners' perceptions of difficulty can lead to "broaden the current understanding of task difficulty and assist language educators in designing and employing more effective language teaching materials" (Tavakoli, 2009a, p.2), but has not yet been very widely employed in TBLT or CAF research. Furthermore, exploring the learners' judgement of TD through both quantitative and qualitative data can be valuable in "assessing task-generated cognitive demands as a way to provide validity evidence for manipulations of task complexity" (Révész et al., 2016, p. 703) and endorsing how participants rate more complex tasks in terms of difficulty (Gilabert & Barón, 2013). Researching TD (e.g., Robinson, 2001; Tavakoli, 2009a; Tavakoli & Skehan, 2005) has showed that learners' perceptions of task difficulty were influenced by task design, and that learners were likely to perceive the more cognitively complex tasks as more difficult. However, the combination of TD perceptions with the fine-

grained linguistic analyses of performance used here was designed to capture a greater level of insight than has been previously shown, and thus add to the validity of existing assumptions.

3.7 Pilot study

A pilot study was designed to a) ensure that the tasks would elicit the predicted language performance; b) fine-tune the instructions of the tasks; c) fine-tune the duration of the video clips; and d) judge the practicality of the data collection procedure. Twelve students participated in piloting the tasks and the questionnaire. All participants were studying in pre-sessional courses at a university in the UK. They speak Arabic as a first language and had been living in the UK for 3-6 months at the time of the pilot study. They reported their IELTS scores as 5.0-6.0. The participants read the task instructions in English and then were assigned to perform the two tasks in a random order. Different durations of each video clips were used, i.e. 60, 90, and 120 seconds. After performing the two tasks, the participants completed the retrospective questionnaire about their perceptions of task difficulty and were further interviewed to get a more in-depth post-task feedback regarding the tasks and procedure of data collection. Each performance was recorded and transcribed.

Piloting the tasks revealed a need to modify the instructions by adding the following statement: *“Imagine that they are narrating the story of each clip to someone who is not watching the video”*. This statement helped creating a genuine purpose to motivate the participants to perform the narrative tasks. The participants also reported that they were confused when they needed to name the characters accurately. They mentioned that trying to remember which was Pat or Mat affected the flow of their narration. Therefore, a colourful picture of Pat and Mat was added to the rubric with a note showing what each character was wearing. The pilot study further revealed a need to give the participants more time after the end of each video clip to finish their speech. A decision was made to include extra 20 seconds in order to enable the participants to complete their narration. A statement about the extra time was therefore added to the task instructions.

As regards the duration of the video clips, 90 seconds were found to be the most ideal option. While the 60-second clip was too short to generate sufficient informative data, the 120-second clip was commonly felt to be overwhelming for the participants to narrate under time pressure. The results of the pilot study further suggested that the selection of the two clips was appropriate for the purpose of the study in eliciting both the predicted oral performance and perceptions of

difficulty. Nine participants out of twelve rated the task that required IR as more difficult than the -IR task. They attributed their perceptions of difficulty to the requirements of the +IR task to read the intentions and thoughts of the characters and to predict and justify their actions while narrating what was happening. Meanwhile, some participants reported that the -IR was easier because the clip included more predictable, logical and familiar actions, and that it required them to only tell and describe what was going on.

3.8 Ethical Procedures

The present research study adhered to Reading University's Ethics Guidance. The ethical procedure was reviewed and approved by the School Ethics Committee. During data collection, the researcher ensured the privacy and confidentiality of the participants, and their right to withdraw from the study at any time. All participants read and signed an information sheet which included detailed description of the research project and a consent form to participate in the study (see Appendix 3). A copy of the signed consent form was given to each participant and another copy was kept in the Department Office of the school.

3.9 Data Collection

The data were collected from the participants individually during a regular school day. The twenty participants were given a brief about the experiment by the school's co-headmaster. A quiet room was prepared by the school administration for the researcher to meet the participants individually. The students were called in an alphabetical order for the meetings. Each one-to-one meeting took about fifteen minutes. The researcher welcomed each participant, introduced himself and first chatted with the student for two minutes to help him relax. After that, the researcher checked the learner's willingness to participate in the experiment and ensured whether he was aware that he had the right to withdraw from the study at any time. The researcher assured the confidentiality of the learner's participation and the anonymity of his identity. Each participant read the information sheet and signed the consent form.

Then, the researcher explained the procedure of the experiment and read the task instructions in L1 and L2. The researcher checked that the participant had understood the requirements of each task. A voice recording software (Audacity, 2012) installed on the researcher's laptop and a high-quality headset were used to record each performance. All the participants performed the two tasks (-IR/+IR) in a counterbalanced order to reduce any effect of practice and order on

performance. The researcher set the recorder and played the video clips (90 seconds each) while the participants narrated the story as they were watching them. If they needed, the participants were allowed extra 20 seconds to sum up their narration at the end of the video clips. After performing the two tasks, each participant completed a questionnaire on his perception of TD, in which he described how difficult each task was and justified his judgement through two open-ended questions.

3.10 Data Coding

The forty recorded speech samples were transcribed using SoundScriber software (Breck, 1998). The data were then coded for selected measures of syntactic complexity, lexical complexity, accuracy and fluency. It is worth mentioning that selecting appropriate measures remains a somewhat challenging area for speech research. Twelve measures had been employed to operationalise the four aspects of learners' language performance. The AS-unit (Foster, Tonkyn, & Wigglesworth, 2000) was used as a main unit for speech segmentation (as discussed in section 3.10.1 below). Inter-rater reliability was checked by recoding 10% of the data by an expert researcher to ensure the reliability and accuracy of the coding process (as discussed in section 3.10.5). A full list of the symbols used in data coding are presented in Appendix 6.

3.10.1 Measures of analysis

Different types of units have been used in L2 research to facilitate the analysis of spoken data. Previous studies of this kind (e.g., Iwashita, 2001; Kuiken & Vedder, 2008; Sangarun, 2005) employed the C-Unit and the T-unit, but these units are now considered as not fit for purpose because they are inadequate to describe the messy reality of spontaneous speech transcripts (Foster et al., 2000). The C-unit which is defined as words or phrases or sentences, grammatical and ungrammatical, which provide referential or pragmatic meaning (Pica, Holliday, Lewis, & Morgenthaler, 1989) does not specify the nature of the independent sub-clausal units. This means that the C-unit does not take into consideration intonational or syntactic units which characterise spoken data (Foster et al., 2000). The other popular unit for the analysis of written and spoken data, i.e. the T-unit, consists of a main clause plus any other clauses which are dependent on it (Hunt, 1970). The T-unit is not suitable either for a full analysis of L2 spoken discourse, which is usually characterized with incidents of dysfluency, reformulations, false starts, hesitations, and repetitions, while it is more sufficient to describe written data. Therefore, it can be argued that the analyst will find it hard to achieve higher levels of reliability by using

C-unit or T-unit because they exclude the speakers' elliptical constructions, resulting in a lot of valuable spoken data being missed (Foster et al., 2000).

Recently, AS-unit has found its way to the heart of task-based research (e.g., Malicka & Levkina, 2012; Qian, 2014; Tavakoli & Foster, 2011) as the basic unit of segmenting spoken data. The AS-unit which is defined as "a single speaker's utterance consisting of either independent clause or sub-clausal unit together with any subordinate clause(s) associated with either" (Foster et al., 2000, p.365) is assumed to be more valid and reliable than the aforementioned units in segmenting and analysing spoken data because it: a) includes any independent sub-clausal units unlike a T-unit; b) it specifies the nature of these sub-clausal units unlike a C-unit; and c) it is better in satisfying the characteristics of L2 spoken data (ibid). Therefore, the AS-unit has been considered as the unit of analysis in this study.

3.10.2 Measures of complexity

Syntactic complexity was operationalised through three measures, i.e. 1) *mean length of AS-unit* (MLASU), calculated by dividing the number of words by the number of AS-units; 2) *mean length of clauses* (MLC), calculated by dividing the number of words by the number of clauses; and 3) *ratio of subordination* (ROS), calculated by dividing the number of clauses by the number of AS-units. Measures of syntactic complexity were designed to capture the length, subordination, subclausal level and coordination of speech production in L2 research (Norris and Ortega, 2009). Following Norris and Ortega (2009), the three measures employed in this study tapped syntactic complexity in terms of a) *length*, which can be measured by any general length-based metric (MLASU in this study); b) *subordination*, which can be measured by subordination measures (ROS in this study); and c) *subclausal level*, which can be measured by clause length-based metric (MLC in this study). However, coordination-based measures of syntactic complexity were not used because they advantage only low language proficiency groups (ibid), while the participants in this study belonged to an upper intermediate level of proficiency. The choice of syntactic complexity measures also considers the recommendations of Inoue (2016) who suggests that "researchers need to consider seriously the task-essentialness of subordinate clauses when deciding on the tasks to use for research" (p.495).

Lexical diversity was used to respond to the linguistic complexity of the participants' speech production. Diversity as a construct refers to the number of diverse linguistic items that can be detected in the learners' language performance (Housen et al., 2012). Previous task-based

studies (e.g. Gilabert, 2007b; Robinson, 2001) advocate that lexical diversity is highly sensitive to the impact of TC. By means of corrected type-token ratios, measures of lexical variety are assumed to tap not only the complexity of the learners' lexical knowledge and proficiency level but also their validity in truly capturing the concept of lexical diversity (Jarvis, 2013).

Previously, a measure of type-token ratio (TTR) was the trend in SLA research to capture lexical complexity of L2 performance. TTR refers to the range of different words (types) used in a text in relation to the total number of words (tokens) (B. Richards, 1987). The problem with the typical TTR measures is their sensitivity to variations in text length. This is solved through the D measure (Malvern & B. Richards, 2002), which uses corrected type-token ratio through a computer program. D values were calculated in this study by using Voc-D which is available in *Coh-Metrix version 3.0 indices* (Graesser, McNamara, & Louwrese, 2003). Before entering the data to Coh-Metrix, the transcribed files were pruned by excluding all the filled pauses and repairs. Voc-D automates the D values by randomly sampling (35-50) words from the transcript to produce a curve of the TTR against tokens for the data (McCarthy & Jarvis, 2010). The software then finds the best fit by adjusting the value of the parameter D. This method is argued to be a valid and reliable measure of lexical diversity that overcomes the problems of variations in sample sizes found in previous methods (Malvern & B. Richards, 2002) . As a result, D is recommended as a measure of lexical diversity for tasks that prompt speech production based on dissimilar contents or themes (De Jong & Vercelloti, 2016).

3.10.3 Measures of accuracy

Two measures of accuracy were used, namely, *percentage of error free clause* (EFC) (Foster & Skehan, 1996) and *errors per 100 words* (EPHW) (Mehnert, 1998). EFC is the number of clauses that contain no errors divided by the total number of independent clauses, sub-clausal units or subordinate clauses in the sample, multiplied by 100. EPHW is the number of errors divided by the number of words produced, multiplied by 100. EFC has been extensively used in a number of studies as a general measure of accuracy because it is predicted to be more sensitive and successful in detecting variation between experimental conditions, and more reliable in dealing with the characteristics L2 data (Skehan & Foster, 1999). Using EPHW is also justified because it is not sensitive to the inconsistency of coding and segmenting clauses, unlike EFC which is dependent on deciding what constitutes a clause (Mehnert, 1998). Furthermore, EPHW as a measure of accuracy is not affected by the spread of the errors and is

argued to be more valid than the unit-based measures of accuracy (Inoue, 2016) . Besides, combining both measures could therefore offer more robust analysis of the accuracy construct.

R. Ellis and Barkhuizen (2005) recommend employing general measures of accuracy, such as percentage of error-free clauses or number of errors per 100 words, but they warn L2 researchers to the need to carefully consider the problems that may arise from the difficulties of determining exactly what constitutes an error, mainly, errors of pronunciation and native-like selection. Given that the phonological features of the participants' language performance (e.g., stress and intonation) were beyond the scope of this study, inaccurate pronunciation that did not affect meaning was not considered as an error in the analysis. As regards errors in native-like selections, a native speaker expert was consulted to resolve any disagreement. It can be argued that the general measures of errors do not typically distinguish between types of errors, i.e. early acquired grammar versus hard to acquire grammar that can be error-prone at even high stages of proficiency (e.g. article choice, 3rd person agreement). However, specify this is beyond the scope of the current study, as the measures are combined with other measures.

In the current study, self-corrected clauses which included reformulations or false starts were counted as error-free. Error-free clauses had no errors in syntax, morphology, word order, native-like selection, or pronunciation that changed meaning. However, errors in stress or intonation were not counted. The British National Corpus (BNC) and Corpus of Contemporary American English (COCA) were frequently consulted when there were doubts about accuracy in terms of native-like selections, in addition to consulting a native speaker expert with respect to any unsolved cases. Samples of error types detected in the data are presented in Appendix 7.

3.10.4 Measures of fluency

Within SLA research, two main sub-dimensions of fluency were targeted in language-based studies. That is, 1) *temporal variables* which are related to the speed of language production, and 2) *hesitation phenomena* which are related to the dysfluency features of performance (Lennon, 1990). In response to this classification, Skehan (2003) recommended using measures of *breakdown fluency* to assess the temporal variables, and *repair fluency* to tap the hesitation phenomena. Following Kahng (2014) and Tavakoli (2011), six measures of fluency were recruited in Study One. That is 1) mean length of mid-clause silent pauses, 2) mean length of end-clause silent pauses, 3) number of mid-clause silent pauses, 4) number of end-clause silent

pauses, 5) number of filled pauses, and 6) number of repairs (repetition, reformulation, and false start).

The decision to use the measures of filled and unfilled pauses was based on previous studies with similar small samples sizes, which found that the frequency of filled and unfilled pauses significantly distinguished between fluent and dysfluent speakers (e.g., Freed, 2000; Lennon, 1990). In addition, this study distinguished between mid-clause and end-clause pauses. The former is argued to be more associated with L2 speakers who are more likely to have more non-functional pauses in the middle of clauses, whereas native speakers tend to use functional silent pauses at clause boundaries to serve a purpose (Tavakoli, 2011). All the measures of fluency were calculated per sixty seconds and the threshold for a pause was set at > 0.25 second following De Jong & Bosker (2013) who “concluded that for the purpose of L2 research, the traditional cut-off point of 250 ms is a good choice” (p. 20). *GoldWave software v5.70* (2009) was used to detect the length of silent pausing in the data.

Repair fluency was represented in this study through the measures of number of repair incidents per minute, which included: 1) *repetitions*: words, phrases, or clauses that are repeated with no modification; 2) *false starts*: utterances that are abandoned before completion; 3) *reformulations*: phrases or clauses that are repeated with some modification; and 4) *replacements*: lexical items that are immediately substituted for another (R. Ellis and Barkhuizen, 2005). As replacement incidents occurred very infrequently in the data, they were counted as reformulations. Samples of repair incidents from the data are shown in Appendix 8. Full descriptions of the twelve measures employed in Study One are presented in Table 4.

Table 4. Measures of complexity, accuracy, and fluency in Study One

| | Dimension | Measure | Abbrev. | Definition |
|----|-----------------------------|-------------------------|----------------|--|
| 1. | Syntactic Complexity | Mean length of AS-units | MLASU | The total number of words divided by the total number of AS-units. |
| 2. | Syntactic Complexity | Mean length of clauses | MLC | The total number of words divided by the total number of clauses. |
| 3. | Syntactic Complexity | Ratio of subordination | ROS | The total number of clauses divided by the total number of AS-units. |
| 4. | Lexical Complexity | Lexical diversity D | D | Adjusted type token ratio computed by Coh-metrix software |

| | | | | |
|-----|-----------------|---|---------------|--|
| 5. | Accuracy | Error-free clause | EFC | The number of error free clauses divided by the total number of independent clauses, sub-clausal units and subordinate clauses multiplied by 100. |
| 6. | Accuracy | Errors per 100 words | EPHW | The number of errors divided by the number of words produced multiplied by 100. |
| 7. | Fluency | Number of mid-clause silent pauses | NMCSP | The total number of mid-clause pauses over 0.25 second divided by the total time of speech in seconds, multiplied by 60. |
| 8. | Fluency | Mean length of mid-clause pauses | MLMCSP | The total length of mid-clause pauses over 0.25 second divided by the total number of mid-clause pauses over 0.25 second. |
| 9. | Fluency | Number of end-clause silent pauses | NECSP | The total number of end-clause pauses over 0.25 second divided by the total time of speech in seconds, multiplied by 60. |
| 10. | Fluency | Mean length of end-clause silent pauses | MLECSP | The total length of end-clause pauses over 0.25 second divided by the total number of end-clause pauses over 0.25 second. |
| 11. | Fluency | Number of filled pauses per minute | NFP | The total number of filled pauses such as <i>eh, er, um</i> divided by the total time of speech in seconds, multiplied by 60. |
| 12. | Fluency | Number of repairs | NR | The total number of repairs such as repetitions, reformulations and false starts divided by the total time of speech in seconds, multiplied by 60. |

3.10.5 Inter-rate reliability

10% of the data was re-coded by an expert researcher independently to check the reliability of the researcher coding for the measures of complexity, accuracy, and fluency. Due to some disagreements in the first round, there was a reassessment of all coded measures by the researcher. The inter-rater reliability coefficients for the measures of complexity and repair fluency were high on the second round, obtaining 91%, whereas the coefficient for the measures of accuracy was still low. To resolve this disagreement, a native speaker expert recoded the accuracy measures for certain types of errors, mainly those involved in inappropriate native-like use. The inter-rater reliability correlation of above 90% was achieved for the accuracy measures in the third round. For samples of the coded transcribed data from Study One, see Appendix 9.

3.11 Data Analysis

The data was computed and analysed using IBM SPSS 21.0. Descriptive and inferential analyses were run for each performance measure in the two tasks (-IR & + IR) to analyse the variance of the participants' performance in terms of complexity, accuracy and fluency. Descriptive analyses (including means and standard deviation) were run to get first impressions about the effects of manipulating Task Complexity (TC) on speech production. The normal distribution of the data was checked before running the inferential analysis (including significance level and effect size). Advanced statistical tests were conducted to figure out if there were any statistical significant differences between learners' performance on the two tasks for each measure.

A repeated-measures (within-participants) Multivariate Analysis of Variation (MANOVA) was conducted to reveal the relationship between the different dependent variables, i.e. measures of CALF in this study. MANOVA was employed because it had the power to detect whether the two performances were different along a combination of measures (Field, 2013). Paired-sample t-tests were then used to detect the variation between the participants' performances in the -IR versus the +IR task, by comparing the means obtained from each measure. Using paired-sample t-test was adequate for the study design which compromised one group in an attempt to compare the performance of the same participant on two different occasions. It is ideal to use paired-sample t-tests when the same participants are exposed to two different experimental conditions (Pallant, 2013). Where significant results were achieved, effect sizes (Cohen, 1988) were calculated and interpreted following (Plonsky & Oswald, 2014). The quantitative and qualitative data of the TD questionnaire were also carefully analysed.

3.12 Conclusion

This chapter provided a detailed overview of the aims of Study One, the research questions and hypotheses, the study design, the participants and the research instruments used in collecting data. The pilot study and the ethical procedure were then acknowledged. The processes of the data collection, transcription, coding and analysis have been discussed, evaluated and justified. The novelty and original contribution of Study One aimed to define and operationalise IR more systematically, and thus, develop a framework for researching IR within the models of TC. The results of Study One will be presented through the descriptive and inferential analyses and answers for the research questions will be offered in the following chapter.

Chapter 4: RESULTS: STUDY ONE

4.1 Introduction

As described in the previous chapter, Study One employed a within-subject design to investigate the effect of TC manipulated by IR demands on L2 learners' oral performance and perceptions of TD. While the independent variable was IR with two levels (-/+), the dependent variables were the four aspects of language performance, i.e. syntactic complexity, lexical complexity, accuracy, and fluency. Twelve measures were used to operationalise the four aspects of L2 oral performance (see Table 4). Perception of TD is another dependent variable in this study which was measured by administering a retrospective questionnaire to explore how the participants perceived the two tasks in terms of difficulty. The twenty participants who performed the two video-based narrative tasks belonged to the same level of language proficiency, i.e. B2 of CEFR.

Descriptive and inferential analyses were run using IBM SPSS 21. Following checking the normality of the data distribution, a repeated-measures multivariate analysis of variance (MANOVA) was run to detect whether there were any statistical significant differences between the performances in the two tasks. Since the results of the MANOVA indicated some statistically meaningful differences, paired-samples t-tests were then conducted to identify where the significant differences were located and answer the research questions. Cohen's *d* (1988) effect sizes were calculated to assess the importance of the significant effects.

4.2 Descriptive analysis

Descriptive analyses including means and standard deviation were run for the twelve measures of oral performance to get a first-impression idea about the impact the independent variable (IR) had on learners' speech production. The results of the descriptive analyses, as presented in Table 5 below, showed that the +IR task elicited speech performance that was characterized with higher accuracy and syntactic complexity, whereas performance under the -IR condition was more lexically complex than the +IR condition. However, the means did not suggest any impact of increased IR demands on the fluency of speech production in the two tasks. The only key differences seemed to be in the measures of repair fluency and filled pausing as the +IR task generated slightly more repair incidents and filled pauses. As regards perceptions of TD, the means revealed that the participants rated the +IR task as more difficult than the -IR task.

Table 5. Descriptive statistics for the dependent variables (Study One)

| Dimensions | Measures | - IR | | + IR | |
|-----------------------------|---|-------|-------|-------|------|
| | | Mean | SD | Mean | SD |
| Syntactic Complexity | Mean length of AS-units | 7.73 | 1.09 | 8.95 | 1.69 |
| | Mean length of clauses | 5.83 | .79 | 5.91 | .81 |
| | Ratio of subordination | 1.33 | .16 | 1.50 | .16 |
| Lexical Complexity | Lexical diversity (D) | 27.37 | 7.36 | 24.62 | 5.91 |
| Accuracy | Percentage of error free clauses | 48.32 | 15.76 | 59.96 | 9.66 |
| | Number of errors per 100 words | 9.36 | 2.77 | 7.53 | 1.73 |
| Fluency | Number of repairs | 6.65 | 4.22 | 7.58 | 5.18 |
| | Number of mid-clause silent pauses | 5.85 | 2.56 | 5.55 | 3.28 |
| | Mean length of mid-clause silent pauses | .79 | .28 | .65 | .26 |
| | Number of end-clause silent pauses | 10.50 | 3.01 | 11.15 | 3.45 |
| | Mean length of end-clause silent pauses | 1.33 | .61 | 1.13 | .46 |
| | Number of filled pauses | 12.27 | 5.28 | 13.5 | 5.11 |
| Task difficulty | Perceptions of Task Difficulty | 2.10 | .55 | 2.75 | .71 |

(*N* = 20)

Prior to conducting any inferential statistical analyses, Kolmogorov-Smirnov and Shapiro-Wilk tests were used to test the assumptions of normality of the data distribution. The results of Kolmogorov-Smirnov and Shapiro-Wilk tests showed that the data were normally distributed with $P > 0.05$. This permitted further advanced inferential statistical analyses.

4.3 Multivariate analysis of variance (MANOVA)

A repeated measures within-group MANOVA was conducted to investigate whether the differences between the two tasks as detected by the preliminary analysis were statistically significant. Pallant (2013) argues that using MANOVA has the advantage of controlling or adjusting the increased risks of a Type 1 error as a result of conducting a series of ANOVAs separately for each dependent variable. Using MANOVA was also justified for its power to look at several dependent variables simultaneously (Field, 2013), and thus detect whether the two performances in this study, were different along the combination of measures.

To avoid reaching misleading conclusions that are statistically meaningful, when in reality, they are likely to be empirically meaningless (Field, 2013) or getting unclear results (Tabachnick & Fidell, 2013) when running MANOVA with many dependent variables (e.g., the 12 measures in this study), it was essential to select representatives of each group of measures for MANOVA. Only four measures were selected as representatives of each dependent variable (CALF) for the MANOVA tests. Additionally, the decision was made to include only four measures because of the small sample size ($N = 20$), and in order not to violate the principles of MANOVA by conducting many comparisons when there were only 20 participants. So with only four comparisons of variation, there would be five participants for each cell which is the minimum for conducting MANOVA (Tabachnick & Fidell, 2013).

The selection was based on the results of factor analyses conducted in previous studies (e.g., Skehan & Foster, 2005; Tavakoli & Skehan, 2005) which concluded that these measures had been distinct factors in TBLT research and the most reliable indicators of the dimensions of L2 performance:

- 1) Syntactic complexity: *Mean length of AS-units (MLASU)*
- 2) Lexical complexity: *Lexical diversity (D)*
- 3) Accuracy: *Percentage of error free clauses (EFC)*
- 4) Fluency: *Mean length of mid-clause silent pauses (MLMCSP)*

The MANOVA analysis on the selected measures showed an overall significant statistical effect of reasoning demands on syntactic complexity (Wilks' Lambda = .632; $F = 11.04$, $p = .004$; $\eta^2 = .368$), lexical complexity (Wilks' Lambda = .654; $F = 10.04$, $p = .005$; $\eta^2 = .346$), and accuracy (Wilks' Lambda = .632; $F = 11.50$, $p = .003$; $\eta^2 = .377$). As for fluency, no significant difference was found between the -IR and +IR tasks (Wilks' Lambda = .866; $F = 2.94$, $p = .102$; $\eta^2 = .134$). It is worth mentioning that the significant difference for lexical diversity was in the opposite direction. The detected significant effects allowed running a number of paired-samples t-tests on all the twelve dependent measures to answer the research questions provided in the previous chapter by identifying where the significant differences were located and estimate the scale of these effects.

4.4 Effects of reasoning demands on oral performance

Paired-samples t-tests were carried out on each dependent variable to compare the means scores of each participant in the two tasks (-IR versus +IR) on different measures of syntactic complexity, lexical complexity, accuracy, and fluency. By employing a repeated-measures design in Study One, it inevitable to address the research questions with paired-samples t-tests to get more statistical power (Field, 2013). With multiple pairs of performances in this study, the paired-samples t-tests checked whether the differences between the means of the pairs are different from zero and whether these differences are statistically significant.

Reaching a significance result does not always give a clear depiction about the importance of the effect (Field, 2013). In order to assess the importance of the findings of this study, the effect size '*Cohen's d*' was calculated, where significant results were achieved (Cohen, 1988). Effect sizes show the relevant proportion of the differences between means, or the amount of variance in the dependent variable that is predictable due to the impact of the independent variables (Tabachnick & Fidell, 2013). *Cohen's d* provides effect size statistics for t-tests by presenting the difference between groups in terms of standard deviation units. Cohen (1988) proposed the following guidelines to interpret the strength of *Cohen's d* effect size statistics: a) below 0.2 is a small effect size; b) 0.5 is medium; and c) 0.8 and above is a large effect size. Applying effect size statistics to this study was a response to the increasing calls for indicating effect sizes in L2 research for their value and usefulness (Norris & Ortega, 2009).

In terms of interpreting the output of the effect size *d*, this study went beyond Cohen's interpretations. Study One adopted the effect size benchmarks suggested by Plonsky and Oswald (2014) who maintained that "Cohen's benchmarks generally underestimate the effects obtained in L2 research (p. 1)". They argued that effect sizes should be interpreted differently within each discipline, and suggested a new set of threshold levels for language-based studies, i.e. a) $d = 0.4$, a small effect size; b) $d = 0.7$, medium; and c) $d = 1.0$, a large effect size.

As shown in Table 6 below, the results of the paired-samples t-tests detected many statistical significant differences ($p < 0.05$) with different effect sizes between the participants' performances in the two narrative tasks (-IR versus +IR). These results are explained in more details in the following sub-sections.

Table 6. Paired-samples t-tests and effect sizes (Study One)

| Dimensions | Measures | - IR | + IR | t-test | Sig | Effect |
|-----------------------------|---|------------------|-----------------|--------------|--------------|-------------|
| | | Mean (SD) | Mean (SD) | <i>t</i> | <i>p</i> | <i>d</i> |
| Syntactic Complexity | Mean length of AS-units | 7.73 (1.09) | 8.95 (1.69) | -3.32 | .004* | .85 |
| | Mean length of clauses | 5.83 (.79) | 5.91 (.81) | -.44 | .659 | .09 |
| | Ratio of subordination | 1.33 (.16) | 1.50 (.16) | -2.96 | .008* | 1.01 |
| Lexical Complexity | Lexical diversity (D) | 27.37 (7.36) | 24.62 (5.91) | 3.17 | .005* | 1.83 |
| Accuracy | Percentage of error free clauses | 48.32 (15.76) | 59.96 (9.66) | -3.39 | .003* | .89 |
| | Number of errors per 100 words | 9.36 (2.77) | 7.53 (1.73) | 2.87 | .010* | .79 |
| Fluency | Number of repairs | 6.65 (4.22) | 7.58 (5.18) | -1.29 | .211 | .19 |
| | Number of mid-clause silent pauses | 5.85 (2.56) | 5.55 (3.28) | .36 | .721 | .01 |
| | Mean length of mid-clause silent pauses | .79 (.28) | .65 (.26) | 1.71 | .102 | .51 |
| | Number of end-clause silent pauses | 10.50 (3.01) | 11.15 (3.45) | -.88 | .389 | .30 |
| | Mean length of end-clause silent pauses | 1.33 (.61) | 1.13 (.46) | 1.71 | .103 | .37 |
| | Number of filled pauses | 12.27 (5.28) | 13.5 (5.11) | -1.70 | .104 | .23 |
| TD | Perception of task difficulty | 2.10 (.55) | 2.75 (.71) | 3.32 | .004* | 1.02 |

Note: -IR = no reasoning required, +IR = reasoning required, *df* = 19, **p* (2-tailed) < 0.05

The outputs of the paired-samples t-tests were used to check the hypotheses that were set to the research questions raised in the previous chapter (see Section 3.3). Study One was designed to answer the following main research question: *Does manipulating TC through the amount of IR required affect learners' L2 oral performance?* This question was divided into four sub-questions and hypotheses to explore the effect of task complexity on syntactic complexity, lexical complexity, accuracy, and fluency of the learners' oral performance.

4.4.1 Effects of IR demands on syntactic complexity

Hypothesis 1a predicted that language performance in the +IR task would be more syntactically complex than the -IR task. Syntactic complexity was operationalised through three measures, namely, *mean length of AS-units*, *mean length of clauses*, and *ratio of subordination*. Participants in the +IR task produced longer AS-units ($M = 8.95$, $SD = 1.69$) than in the -IR task ($M = 7.73$, $SD = 1.09$). A statistical significant difference was detected in terms of *mean length of AS-units* ($t = -3.323$, $p = .004$) with a medium effect size ($d = .85$). Moreover, a higher *ratio of subordination* was marked in the +IR task ($M = 1.50$, $SD = .16$) than in the -IR task ($M = 1.33$, $SD = .16$). The difference between the two tasks was statistically significant with a large effect size ($t = -2.962$, $p = .008$, $d = 1.01$). As regards *mean length of clauses*, the +IR task produced longer clauses ($M = 5.91$, $SD = .81$) compared to ($M = 5.83$, $SD = .79$) in the -IR version. However, the variation was not statistically significant ($t = -.449$, $p = .659$). The results of the t-tests revealed that the independent variable in this study, i.e. IR had a positive effect on syntactic complexity of the learner's language performance in terms of *mean length of AS-units* and *ratio of subordination*, but not for *mean length of clauses*. Therefore, *Hypothesis 1a* was partially confirmed through the statistical significant effects of TC on two measures of syntactic complexity out of three.

4.4.2 Effects of IR demands on lexical complexity

Hypothesis 1b predicted that oral language performance in the +IR task would be associated with more lexical complex language than in the -IR task, measured through lexical diversity, i.e. D. However, the results pointed in the opposite direction of the hypothesis. It was in fact the -IR task that generated more lexically varied language ($M = 27.37$, $SD = 7.36$) than the +IR task ($M = 24.62$, $SD = 5.91$). This difference was statistically significant with a large effect size ($t = 3.170$, $p = .005$, $d = 1.83$). The findings thus refute *hypothesis 1b*. It seems, counter to predictions, that increasing TC through IR demands had a negative effect on L2 lexical performance in this study.

In order to ensure a thorough understanding of the results of lexical diversity, a post-hoc lexical frequency analysis was conducted using Compleat Lexical Tutor (Cobb & Free, 2015) to categorise the lexical items used more frequently in the +IR versus -IR conditions. A summary of the lexical frequency is presented in Table 8 below.

Table 7. Lexical analysis of +IR vs -IR tasks (Study One)

| <i>Word types</i> | <i>Number of times (+IR)</i> | <i>Number of times (-IR)</i> |
|-------------------------------|------------------------------|------------------------------|
| Mental state verbs | 152 | 41 |
| Conjunctions | 124 | 43 |
| Modals verbs | 99 | 27 |
| Adverbs of uncertainty | 42 | 5 |
| Total | 417 | 116 |

As Table 8 showed, the results revealed that mental state verbs (*e.g., think, wish, assume, seem, sound, want, etc.*) were the most frequent group of words used in the +IR condition compared to the -IR condition. The use of logical cohesive conjunctions (*e.g., so, but, because, etc.*) were the second most frequent word types needed in the +IR task. Using more conjunctions in the more complex task could also explain why performance in the +IR task was more syntactically complex than -IR task. Modal verbs (*e.g., may, might, can, could, will*) were also used more frequently in +IR performance, but they were used infrequently in the -IR task. The last category was the adverbs of uncertainty (*e.g., maybe, apparently, perhaps, and probably*), which were used 42 times in the +IR task compared to only 5 times in the -IR task.

4.4.3 Effects of IR demands on accuracy

Hypothesis 1c predicted that language performance in the more complex task, i.e. +IR would be more accurate than performance in the less complex one (-IR). Two measures were employed to operationalise accuracy in this study, i.e., *percentage of error free clauses* and *errors per 100 words*. The +IR task significantly produced more error-free clauses ($M = 59.96$, $SD = 9.66$) compared to the -IR task ($M = 48.32$, $SD = 15.76$). The difference was statistically significant ($t = -3.392$, $p = .003$) with a medium effect size ($d = .89$). A similar result was obtained for *errors per 100 words*, as the participants in the +IR task committed less errors per 100 words ($M = 7.53$, $SD = 1.73$) compared to ($M = 9.36$, $SD = 2.77$) in the -IR task. The difference also achieved a statistically significant level ($t = 2.878$, $p = .010$) with a medium effect size ($d = .79$). The findings obtained from the t-tests revealed that IR demands had a systematic positive impact on the accuracy of the learner's L2 speech production. Hence, it can be concluded that *Hypothesis 1c* was broadly confirmed since performance in the more complex task (+IR) was clearly more accurate than performance in the less complex task (-IR).

4.4.4 Effects of IR demands on fluency

Hypothesis 1d predicted that the participants' oral language performance in the +IR task would be less fluent than the -IR task. Six measures were employed to operationalize fluency in Study One, i.e. *number of repairs, number of mid-clause silent pauses, mean length of mid-clause silent pause, number of end-clause silent pauses, mean length of end-clause silent pauses, and number of filled pauses*. All the measures of fluency were calculated per one minute. In terms of *number of repairs*, the participants in the +IR task produced more repairs incidents ($M = 7.58, SD = 5.18$) than in the -IR task ($M = 6.65, SD = 4.22$), but the difference did not reach a statistically significant level ($t = -1.295, p = .211$). The same results were found with regard to *number of filled pauses* as the +IR task generated a higher number of filled pauses ($M = 13.5, SD = 5.11$) compared to the -IR task ($M = 12.27, SD = 5.28$). Again, the variation between the two tasks failed to achieve a statistical significant level ($t = -1.706, p = .104$). Mixed results were achieved with respect to the measures of silent pausing which failed to reach statistical significant levels. The mixed and non-significant results of fluency measures confirmed that manipulating TC by IR requirements had no effects on the fluency of learners' oral performance, and therefore *Hypothesis 1d* was not confirmed.

4.5 Effects of IR on perceptions of TD

A retrospective questionnaire on how participants perceived the tasks in terms of their difficulty was used to answer Research Question 2: *Do L2 learners perceive the more complex task as more difficult?* Following this question, it was hypothesized that L2 learners would perceive the task that required IR as more difficult than the -IR task. The questionnaire included two scale-questions with four adjectives that described each task ($1=very\ easy, 2=easy, 3=difficult, 4=very\ difficult$) and two open-ended questions to explain what tasks aspects affected the participants' ratings. The questionnaire enabled the study to collect quantitative and qualitative data to get more in-depth insights about learners' perceptions of TD.

As shown in Figure 5 below, the quantitative results of the TD questionnaire revealed that most of the participants perceived the +IR task as more difficult ($M = 2.75, SD = .71$) than the simple task ($M = 2.10, SD = .55$). The output of the paired-samples t-test detected a statistical significant difference between the participants' perceptions of TD in the two tasks ($t = 3.32, p = .004$) with a large effect size ($d = 1.02$). Consequently, *hypothesis 2*, i.e. L2 learners would perceive the more cognitively complex task as more difficult was confirmed.

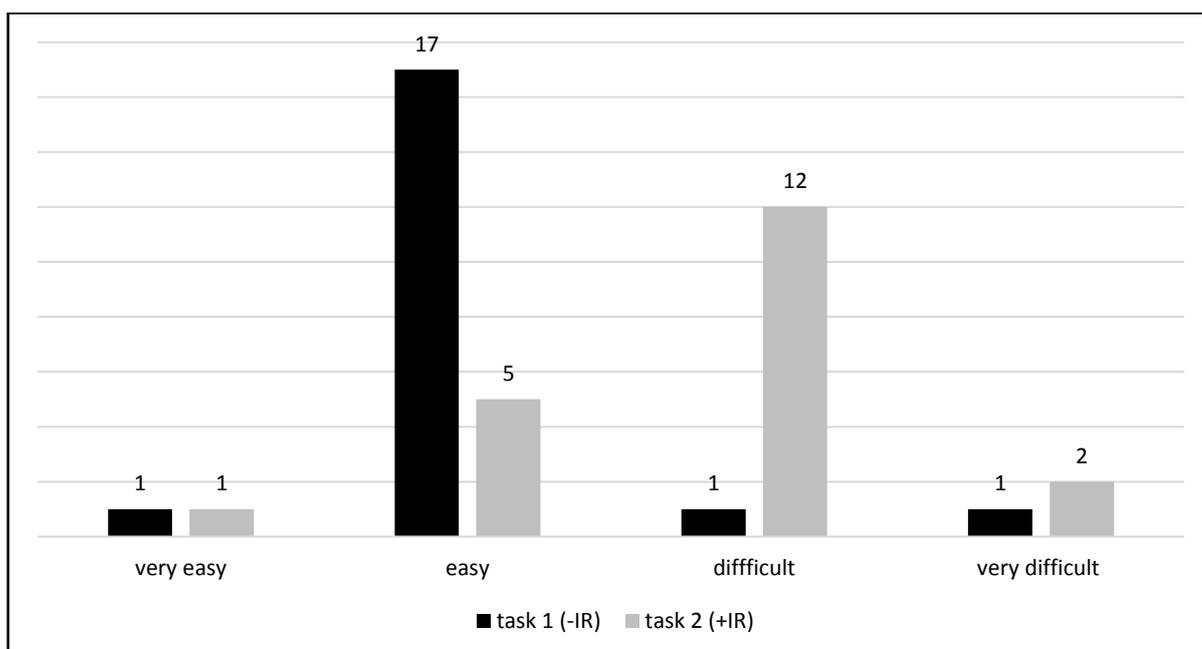


Figure 5. Participants' perceptions of Task Difficulty (Study One)

Looking at the thematic analysis of the questionnaire, the qualitative data revealed a number of factors that affected the participants' judgement with respect to perceptions of TD. The most important emerging theme was the greater cognitive demands accompanying the +IR condition, i.e. the requirement to explain, reason, and predict the characters' intentions and actions while speaking and describing the events. The participants' responses about the cognitive demands were classified into two groups. The first group were statements in which the participants referred to task-inherent cognitive demands, primarily to do with topic familiarity (43% of the responses). For example, they found the +IR task to be more difficult since the video contained unpredictable or less familiar events which were more difficult to understand and predict. One of the participants wrote: *"It is difficult because that's not common in normal life. So I needed some thinking about it."* The participants likewise described the actions in the -IR video as more common and predictable. A participant rated the -IR task as easier, noting that: *"It is from our everyday life and I am used to these kind of things."*

The second group were statements related to task-induced cognitive demands, i.e. the demands caused by the explicit task instructions that encouraged the participants to provide reasoning about the characters' intentions, predicting what they would be doing next, and justifying their actions (41% of the responses). One of the participants contended: *"It is a bit difficult because you have to be ahead of the events."* Another noted: *"I found it difficult because I have to read their minds and what they are thinking about."*

Two other themes emerged but these were much less frequent. Linguistic demands including the need for a lexical item or a specific structure required to narrate the story was also a common theme characterising their perceptions of difficulty in how well they could express the complex reasoning required. A participant who rated the -IR task easy explained: *“The task doesn’t need any hard words and meanings.”* The last theme that emerged from the thematic analysis was the effects of time pressure on language performance and meeting the tasks requirements. One of the participants who found the +IR task more difficult mentioned that: *“The story is live and I have to think so fast to describe the actions.”* It is worth noting that the two tasks were equal in terms of speaking in real time. However, most of the comments that referred to the effects of time were made with respect to the +IR task. The main emerging themes from this qualitative analysis, i.e. task-inherent and task-induced demands will be linked with the key models of task complexity and task difficulty, and will be exposed to a further discussion in the Discussion Chapter. Table 07 below summarises the main emerging themes from the participants’ answers on the two open-ended questions, their frequency and percentages.

Table 8. Thematic analysis of perceptions of TD (Study One)

| Themes | Frequency | Percentage |
|--|------------------|-------------------|
| Cognitive demands (task-inherent) | 23 | 43% |
| Cognitive demands (task-induced) | 22 | 41% |
| Linguistic demands | 5 | 9% |
| Time pressure | 4 | 7% |

Total number of comments: 54

4.6 Summary of key findings (Study One)

This study was mainly designed to investigate the effects of Task Complexity (TC) manipulated by intentional reasoning (IR) demands on L2 oral performance and perceptions of Task Difficulty (TD), and sought, in a novel and original way to address the unsystematic IR definitions and operationalisation proposed by previous research. The study hypothesized that increasing TC through IR would result in positive gains in terms of syntactic complexity, lexical complexity, and accuracy, whereas, negative gains were predicted with respect to fluency. The findings of Study One confirmed the positive effects of IR on the syntactic complexity and accuracy of learners' oral performance, whereas the significant effects on lexical complexity were in the opposite direction in favour of -IR condition. The results further showed that IR demands had no effects on the fluency of oral language performance. As for the data obtained from the questionnaire, the participants perceived the +IR task as more difficult than the -IR task, which confirmed what the study predicted. The participants mainly attributed their perceptions of TD to task-inherent cognitive demands triggered by task content and task-induced cognitive demands triggered by the instructions of each task. All the findings presented in the Results Chapter will be discussed in more detail in the next chapter, finishing with implications for the research goals and design of a follow-up experiment aiming to confirm the findings of the current study and examine the other possible variables that may interact with the effects of increasing TC through IR demands.

Chapter 5: DISCUSSION: STUDY ONE

5.1 Introduction

This chapter discusses the findings of Study One in relation to the research questions and the hypotheses. The discussion chapter will evaluate the current results with respect to the findings of previous studies of TC. The effectiveness of the measures employed in this study will be evaluated. This chapter will then explore the relationship between IR, TC, and TD. The chapter concludes with a discussion of the limitations of the current study as well as implications for designing a follow-up second study, along with wider suggestions for future research beyond the scope of this project.

5.2 Overview of Key findings

The main purpose of the current study was to explore the effects of manipulating cognitive task complexity on L2 oral performance and perceptions of TD. The study further attempted to offer a more systematic operationalisation of IR and explore its relationship with TC and TD. Study One drew on quantitative and qualitative data gathered from 20 secondary school students studying English as a foreign language in Jordan. The mixed-method study design included performing two video-based oral narrative tasks with two levels of TC and completing a retrospective questionnaire on TD.

Following the predictions of the Cognition Hypothesis (CH) (Robinson, 2001, 2005, 2007) that increasing TC through IR demands would be associated with increased syntactic complexity, lexical complexity, accuracy, and decreased fluency, the findings revealed that the +IR task produced more syntactically complex and accurate speech. However, +IR demands resulted in less lexical complexity and inconsistent performance with respect to fluency. Furthermore, the task with +IR requirements resulted in higher perceptions of TD.

In more detail, +IR produced more syntactic complexity measured by *ratio of subordination* and *length of AS-units*, and accuracy measured by *error-free clauses* and *errors per 100 words*. Contrary to what this study hypothesized about the positive effect of IR on lexical complexity measured by *D*, the results showed that more lexically complex language was produced in the less complex task (-IR). Moreover, the predicted negative influence of IR on the fluency of learners' oral performance was not confirmed for measures of *filled and unfilled pauses* and

number of repairs. In terms of TD, the task with +IR requirements was rated as more difficult as the CH predicted. The results showed that the high cognitive demands of +IR imposed through both task content and task instructions of the +IR task were the prevailing contributors to the participants' perceptions of difficulty. In the following sections, the key findings will be further discussed with respect to the effects of IR demands on each aspect of oral performance and the interaction between the results and the choice of measures of analysis. It worth noting that the data presented in the Results Chapter of Study One will be reproduced in forms of figures and charts and included here in the Discussion Chapter for ease of reference and to serve as visual support during discussion.

5.3 Intentional reasoning and syntactic complexity

Research Question 1a asked whether (-/+) IR demands would affect the syntactic complexity of L2 oral performance. It was predicted, following the CH, that speech production in the more complex task (+IR) would be characterized with more syntactic complexity than its less complex counterpart (-IR). The results confirmed the predicted positive influence of IR demands on syntactic complexity with regards to *mean length of AS-unit* and *ratio of subordination* but not *mean length of clauses*. It appeared that the requirements of IR which were operationalised through both task instructions and task content encouraged the participants to produce longer AS-units and use more subordination at the expense of length of clauses.

The reported findings regarding syntactic complexity were consistent with the only study that investigated the effects of IR on L2 oral performance using monologic oral tasks, i.e. Ishikawa (2008), who found that IR demands promoted syntactic complexity in terms of *S-Nodes per T-units*. However, the results of the current study contradicted Robinson (2007) who found that IR had no effect on syntactic complexity in terms of *clauses per C-unit*. It is worth mentioning that Robinson (2007) employed dialogic tasks, but not monologic, to elicit speech production. As a result, the lower grammatical complexity in the more complex tasks in Robinson's study could be due to the effect of interaction and turn-taking between participants which might have reduced the number of clauses per C-unit. On the contrary, the current study, which employed monologic tasks, did not require any interaction or negotiation for meaning.

The results of syntactic complexity did not support those reported by Levkina & Gilabert (2012) who increased TC along (+/*few elements*), and Malicka (2014) who investigated the effects of -/+*reasoning* and +/-*few elements* on L2 performance. The two studies found that manipulating

TC had no effect on syntactic complexity measured by *ratio of subordination*. The results of Study One were also incompatible with the TC research meta-analysis of Jackson and Suethanapornkul (2013) who found negative but negligible effects of TC on syntactic complexity. Figure 6 below shows the difference in means of syntactic complexity measures, i.e. *mean length of AS-unit* (MLASU), *mean length of clauses* (MLC) and *ratio of subordination* (ROS) between the two tasks.

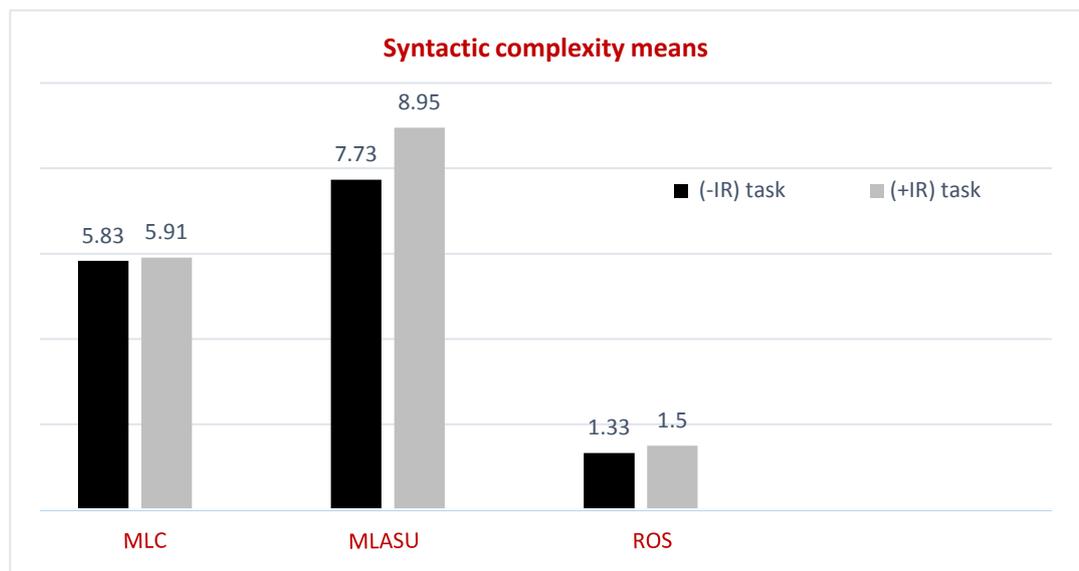


Figure 6. Means of syntactic complexity performance (Study One)

5.4 Intentional reasoning and lexical complexity

Research Question 1b tested the impact of (-/+ IR) on the lexical complexity of oral performance. The hypothesis was formed in line with the Cognition Hypothesis, which anticipated a rise in lexical complexity when cognitive complexity of a task is increased through reasoning demands. The findings were in stark contrast to what had been predicted. It was the less complex task (-IR), which elicited more lexically diverse speech production than the complex counterpart (+IR), measured by D. The difference was statistically significant with a large effect size ($d = 1.83$), but in the reverse direction to the predictions, and in favour of the less complex task. These results challenge the assumption from the Cognition Hypothesis that oral language performance in more cognitively complex tasks is assumed to be more lexically complex. However, it was found that there might be a task-effect on specific language usage about predicting intentions, entailing a narrower lexical range, which will be discussed more fully

below. Figure 7 below demonstrates the difference between the means of lexical complexity, i.e. *corrected type token ratio* (D) across the two tasks in Study One.

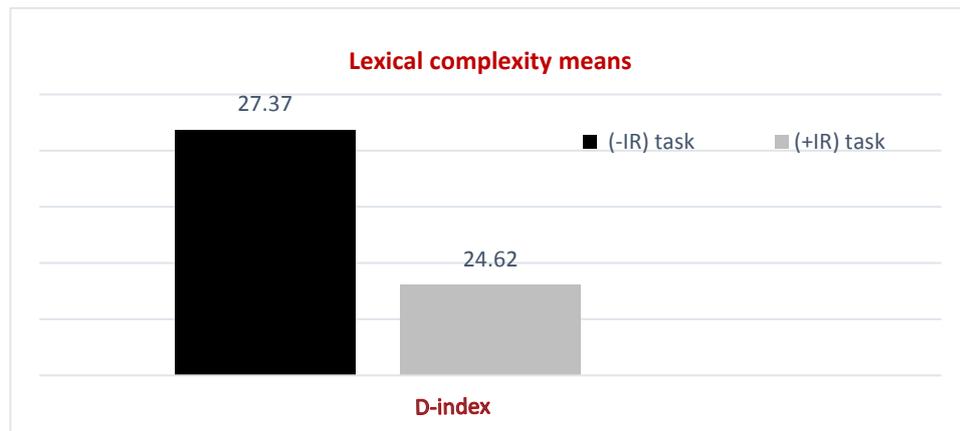


Figure 7. Means of lexical complexity performance (Study One)

These results confirmed those of Robinson (2007) who found that the tasks that required IR generated lower lexical complexity measured by TTR. However, using specific measures, i.e. *psychological and cognitive state verbs*, Robinson found that +IR tasks produced more complex lexis. The findings of the current study contradicted those obtained by Ishikawa (2008) who discovered that tasks with increased IR demands enhanced lexical complexity measured by *Guiraud's index* (2000). Moreover, Malicka (2014) who manipulated TC through -/+ few elements and -/+ reasoning demands using three monologic problem-solving tasks, proposed that TC demands enhanced lexical complexity, measured by *D* and *Guiraud's index*. Jackson and Suethanapornkul (2013) further revealed in their meta-analysis that TC had a positive but negligible influence on lexical complexity measured by general measures.

Levkina and Gilabert (2012) found that the more complex tasks along a resource-directing variable, i.e. + *number of elements*, and also along a resource-dispersing variable, i.e. - *pre-task planning*, elicited more lexically complex but less fluent performance, whereas accuracy and syntactic complexity remained unaffected. While the findings of Levkina & Gilabert supported CH with regard to lexical complexity and fluency, the current study questioned the viability of the CH predictions about lexical diversity. The results of lexis in this study were consistent with Albert (2011) who confirmed that the more cognitively complex task, i.e. *inventing a story*, produced less lexically diverse performance than the simple task, i.e. *telling a given story*. Albert employed measures of *D* and *Plex Lambda* to operationalise lexical complexity.

The less lexically diverse performance captured in the more complex task can be justified as an indication that the task with greater cognitive demands was not successful in directing the learners' attention to form-meaning association. Hence, increasing cognitive task complexity through greater reasoning demands did not seem to help learners to spare more attentional resources to their linguistic performance. A further possibility is that the learners in this study had prioritised syntactic complexity on producing more lexically complex language. This is one of the many possible explanations which cannot be confirmed without further systematic research examining it. The question why accuracy and syntactic complexity improved but not lexical complexity could be answered through the participants' qualitative recall comments, who might have decided not to take any risk in using more varied language, but instead they used more conservative language to focus on the increased cognitive demands of the +IR task.

The increased lexical diversity in the -IR task could be seen as a domino effect of the number of different events the participants had to narrate, which required higher ratio of word types against tokens. On the other hand, the +IR task required a lower ratio of word types because learners were reasoning about the ability of the characters to make their car fly, which meant that the participants were addressing the same issue repeatedly, resulting in more tokens of the same words. Therefore, the linguistic requirements of the +IR task imposed by task instructions and task content stimulated the participants to use more formulaic chunks (e.g. *it seems, it looks like, it could be, I assume, I think, etc.*) to explain the characters' intentions resulting in more repetitive language performance which reduced the D index in the +IR task. This conclusion was further confirmed by the qualitative analysis of the participants' lexis in the two tasks.

The results of this lexical analysis offered some initial evidence to explain the reduced lexical complexity captured by D in the +IR task compared to the -IR task, a result that contradicted the CH predictions. These findings supported the assumption that the content of task is an important factor in shaping the participants' linguistic performance with respect to lexical complexity, and hence selecting the measurement of lexical complexity more carefully. This conclusion was in line with Tavakoli and Foster (2008) and Foster and Tavakoli (2009) who found no effects of increasing TC on lexical diversity measured by D. Their interpretation was that the content of the task, rather than its complexity, drove lexis forward.

5.5 Intentional reasoning and accuracy

Research Question 1c asked whether (-/+) IR demands would affect the accuracy of L2 learners' oral performance in terms of *percentage of error-free clauses*, and *number of errors per 100 words*. The CH predicted that performing oral narrative tasks with +IR requirements would be associated with more accurate language than performing tasks with no IR demands. The results obtained for the measures of accuracy employed in this study clearly supported the claims of the CH. The more complex task (+IR) significantly encouraged more error free clauses and less errors per 100 words compared to the less complex task (-IR). The size of the positive effect size was large for each of the accuracy measures.

Similar results with regards to accuracy were reported in other studies (e.g., Albert, 2011; Ishikawa, 2008; Kuiken & Vedder, 2011; Malicka, 2014; Révész, 2011; Robinson, 1995, 2001) which detected positive effects of performing more cognitively complex tasks on the accuracy of L2 oral performance measured by general and specific measures. Furthermore, the findings of the current study were in harmony with Jackson and Suethanapornkul's (2013) meta-analysis which found a positive impact (though small effect size) of manipulating TC on the accuracy of L2 speech production. However, the results of accuracy measures in this study were not in harmony with Robinson's (2007) findings which suggested that IR demands on dialogic narrative tasks had no impact on accuracy measured by *percentage of error-free C-units*.

The hypothesised explanation of achieving positive results for accuracy is that while reasoning about people's intentions and thoughts in the complex task, the participants would have been encouraged to focus on form (Swain, 1995), resulting in producing fewer errors and more error-free clauses. In the -IR task, where learners had only to tell and describe a series of events, their attention was directed to convey meaning rather than form. This could be a reason why the less complex task (-IR) elicited less accurate language. It can be also argued that there is a close connection between lexis and grammatical accuracy. In this study, it was found that the IR requirements had a direct influence on accuracy due to the extensive use of formulaic language, e.g. *I think, I suppose*, which was recycled to depict the intentions of the characters, leading to AS-units with shorter clauses. As a result, accuracy enhanced dramatically in the +IR condition. However, there is still a need for more in-depth examinations to confirm the link between improved accuracy and the formulaic nature of language used to perform a task. Figure 8 below

compares the means of accuracy measures, i.e. *percentage of error-free clauses* (EFC), and *number of errors per 100 words* (EPHW) in Study One.

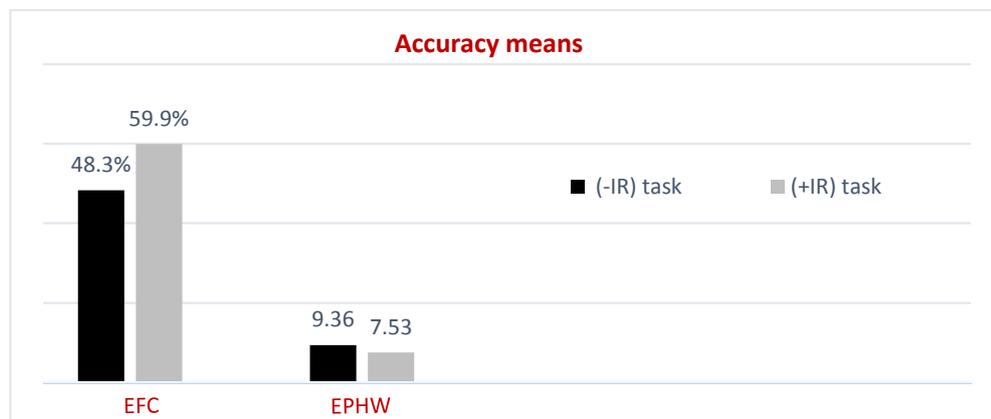


Figure 8. Means of accuracy performance (Study One)

5.6 Intentional reasoning and fluency

Research Question 1d looked into the impact of cognitive task complexity on the fluency of the learners' L2 performance. It was anticipated that the oral language performance on the more complex task would be characterized with less fluent production than on the simple version as presumed by the Cognition Hypothesis. Six measures were employed to operationalize fluency in this study, i.e. *number of repairs*, *number of filled pauses*, *number of mid-clause silent pauses*, *mean length of mid-clause silent pauses*, *number of end-clause silent pauses*, and *mean length of end-clause silent pauses*. All the measures of fluency were calculated per minute. The results failed to confirm the predicted negative effect of performing complex tasks with increased reasoning demands on fluency. Regardless of the less fluent performance on the more complex task in terms of the *number of repairs* and *filled pauses* (non-significant), the four measures of silent pauses showed mixed results on the two tasks. These findings indicated that the independent variable, i.e. IR demands, had no effect on the fluency of learners' oral performance. The results regarding fluency confirmed the findings of (Gilabert, Barón, and Levkina (2011); Michel (2011); Robinson (2007)) who reported no effect of manipulating cognitive task complexity on the fluency of L2 speech production. However, manipulating TC in other studies (e.g., Albert, 2014; Ishikawa, 2008; Malicka, 2014) revealed reverse effects on fluency in terms of *speech rate* and *repair fluency*, offering support to the CH. These results were also confirmed by Jackson and Suethanapornkul (2013) who in their meta-analysis detected a negative impact with small effect size of increasing TC on the fluency of L2 learners' performance.

It should perhaps be noted that while the participants in the current study were performing the more complex task, they seemed to use more filled pauses to buy time in order to meet the required reasoning demands, whereas their silent pausing behaviour was stable across the two tasks. This steady performance in terms of silent pausing could be due to the fact that performing monologic tasks does not seem to encourage L2 speakers to attend to fluency in the same way performing interactive tasks does. It is also possible that the threshold used for a silent pause in this study, i.e. 0.25 second, was not short enough in detecting differences in the pausing phenomena between the less and more complex tasks. The learners' level of proficiency (B2) could be another possible explanation why learners did not pause more on the complex task. Their proficiency level might have diminished the negative effect of increasing cognitive demands on fluency (Tavakoli & Skehan, 2005). The learners might have prioritized using filled pauses over using silent pauses as a more effective strategy to buy time and disguise their dysfluency. The unexpected results of fluency could also be attributed to the fact that the differences between the demands of the -IR task versus +IR task were not sufficient enough to cause problems for participants to maintain their fluency.

5.7 Intentional reasoning and Task Difficulty

Answers to Research Question 2 about how the participants perceived the tasks in terms of their difficulty were obtained through a retrospective questionnaire. The Cognition Hypothesis predicted that L2 learners would perceive the more cognitively complex tasks as more difficult than the simple versions. The current study assumed that the subjective perceptions of task difficulty will match detailed objective measures of linguistic performance – one of the novel methodological elements of this study. The t-tests results showed that the more complex task (+IR) was rated significantly more difficult than the less complex version (-IR). The findings confirmed what was hypothesized in this study and offered empirical support to the Cognition Hypothesis. These findings were further compatible with the majority of previous studies investigating the effects of cognitive task complexity on the perception of task difficulty (e.g., Ahmadian, 2012; Gilabert, 2007b; Préfontaine, 2013; Révész et al., 2016; Robinson, 2001; Tavakoli, 2009a; Tavakoli & Skehan, 2005). In addition, the results with respect to TD confirmed what was reported in similar studies which employed increased IR demands in the complex tasks (Ishikawa, 2011; Robinson, 2007). This could lead to confirm that IR demands have a clear and direct impact on learners' perceptions of TD, and therefore needs to be considered as a key element when designing and classifying tasks. The steady relationship

between TC and TD is assumed to advance the attempts to design a reliable framework to assess TD more systematically. Figure 9 below, depicts a comparison between the perceptions of the two tasks (-IR versus +IR) in terms of their levels of difficulty.

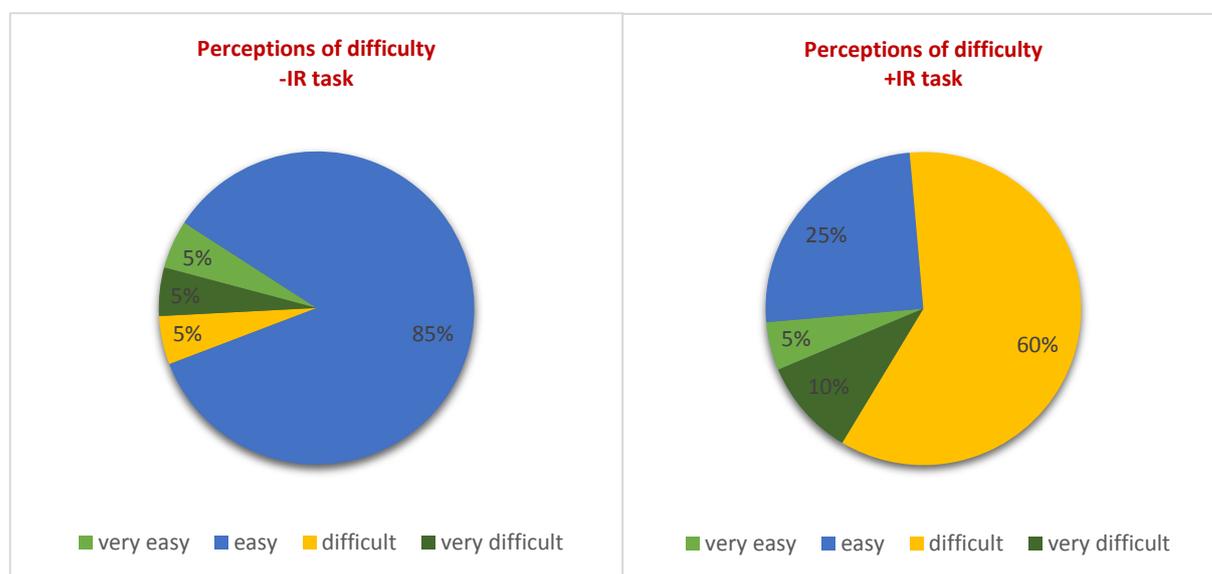


Figure 9. Percentage of Task Difficulty rating (Study One)

Rating the more complex task as more difficult in this study could be regarded as a validation of the effective choice of tasks in terms of manipulating TC. This was in harmony with Révész et al. (2016) who argued that TD self-judgement would help researchers “establish the validity of cognitive task complexity manipulations, that is, to provide independent evidence that tasks that were designed to be more complex do, in fact, place greater cognitive demands on research participants” (p.733). Furthermore, the results of the careful analysis of the qualitative data elicited from the TD questionnaire helped to further validate the two-level framework which this study suggested to operationalise IR more systematically. Exploring TD in this study as a dependent variable and identifying the factors that were linked to the learners’ perceptions helped to depict a richer picture regarding the relationship between task factors and the participants’ factors. However, what was not depicted in this study was investigating the link between the language performance of each task and how the performers perceived the task with respect to difficulty. This could have been possible by employing TD as an independent variable and trying to find out whether certain linguistic patterns can be linked to the participants’ perceptions of TD. However, this was beyond the main scope of this study.

Operationalising TC through IR demands at two levels in this study, i.e. *task content* and *task instruction* was confirmed through the participants' justification of their self-ratings of TD. The participants attributed their perceptions of difficulty to the +IR task cognitive demands imposed primarily by the content of the +IR video clip (task-inherent demands), and task instructions (task-induced demand). The main two themes that were revealed by the qualitative analysis verified; a) the successful selection of the clip for the +IR task as its content drove the task performers to higher levels of cognitive processing, i.e. reasoning, predicting and justifying ongoing actions; b) the effectiveness of +IR task instructions which encouraged the participants to associate their narration with higher levels of cognitive demands requiring them to read the characters' minds and intentions. Combining the linguistic demands (tell and describe) with the cognitive demands (reasoning about others' intentions) is also assumed to affect each aspect of L2 speech production differently. Consequently, the impact of manipulating TC as suggested in the current study can be better interpreted and understood by linking it with the choice of CALF measures of analysis in this study, which I now discuss in more detail to highlight the value of the rich methodological framework used here.

5.8 Intentional reasoning and CALF measures

The measurement-related issues are at the heart of any discussion of the effects of manipulating task designs and conditions on L2 speech production (Lambert & Kormos, 2014). The validity, reliability, and feasibility of the measurements of analysis have been questioned in a bulk of task-based studies (Housen et al., 2012). The findings of Study One offered information on how manipulating TC through IR demands interacted with the selected measures to affect the way learners' oral performance was captured in terms of syntactic complexity, lexis, accuracy, and fluency. The results of this study revealed that the linguistic requirements imposed by IR demands (e.g., *reading thoughts, justifying and predicting actions*) can be captured differently based on the choice of the measurement of analysis.

Measures of syntactic complexity appeared to be influenced by the language required in the +IR task. Whilst two measures of syntactic complexity, i.e. *mean length of AS-unit* and *ratio of subordination* enhanced significantly in the +IR task, the difference between the two tasks was not significant with respect to *mean length of clause*. Expectedly, expressing intentionality and justifying the character's actions were linked with using more subordination (e.g., *I assume they are trying to make their car lighter to fly*). As a result, the excessive use of subordination

resulted in longer AS-units, but shorter clauses. Apparently, the constant use of formulaic chunks to hypothesise about the characters' thoughts, intentions, and actions in the +IR condition triggered using many short clauses (e.g., *It seems that, I think, I suppose, I wonder*) almost in each AS-unit. Consequently, ratio of subordination can superficially be conflated by adding short clauses such as *I think* and *I suppose*. It can be concluded that the increased IR demands complexified the learners' language performance with respect to subordination and length of AS-units; this in turn, appeared to prompt using short clauses more frequently. Therefore, it can be recommended that results of measuring *mean length of clause* should be interpreted cautiously if researchers need to offer reliable perspectives to syntactic complexity. These findings would bring attention to the debate regarding the vital effects of the choice of the measures of analysis on the obtained results. This conclusion is in harmony with Inoue (2016) who suggests that the measures of syntactic complexity should be carefully selected based on "the differing degrees of task-essentialness for subordination between the two tasks" (p.487). Moreover, these results reveal a need to employ more refined measures of syntactic complexity as the traditional length-based and count-based measures of clauses can be ineffective in detecting the different facets of syntactic complexity (Norris & Ortega, 2009).

Measuring lexical complexity using D, showed that the -IR condition encouraged more diverse language. As mentioned earlier, it appeared that the cognitive task complexity imposed by the instructions and content of the +IR task encouraged more formulaic and repetitive patterns of lexis which contributed to the lower D values in the more complex task. These results supported Jarvis (2013) who called for including measures of size, richness, dispersion and sophistication to gain more in-depth investigations of lexical complexity. Therefore, it can be concluded that employing only one measure to capture lexis performance will affect the generalisability of the findings. The post hoc frequency analysis of lexis confirmed the assumptions that the inherent and induced cognitive task demands directly affected the diversity of the participants' language. These demands urged the repetitiveness of mental state verbs, modal verbs, conjunctions, and adverbs of uncertainty resulting in less lexical complex oral performance in the +IR condition which contradicted the predictions of the Cognition Hypothesis.

Turning now to discuss the interaction between IR demands and measures of accuracy, it was found that the two measures employed in the study, i.e. *error-free clauses* and *errors per 100 words* were effective in capturing the variation between the two performances in terms of accuracy. The findings revealed that the +IR condition resulted in significantly more accurate

language with large effect sizes. Again, these results can be linked to the use of repetitive formulaic sequences which resulted in shorter clauses. These patterns helped the participants to produce more error-free clauses and commit less errors per 100 words compared to the -IR condition. Although, the two measures showed more accurate performance in the +IR task, the effect size was not the same for each measure. The effect size for *percentage of error-free clauses* was larger ($d = .89$) compared to *errors per 100 words* ($d = .79$). This difference indicated that each accuracy measure produced slightly different results. This will nurture the debate on “how the two accuracy variables could reveal such different results, and which variable better reflects the actual differences in accuracy of the performances” (Inoue, 2016, p.498).

Similar to the findings of this study, Inoue (2016) found that the reason for the variation between the two measures of accuracy was due to the differences in the denominators, i.e. 100 words vs clauses. As a result, the frequency of the errors and their distributions can contribute directly to the different results, in addition to how accurate clauses are coded. In other words, if two participants have the same percentage of error-free clauses, they might have different number of errors per 100 words. Inoue (2016) argues that errors per 100 words as a measure of accuracy is more valid and sensitive than error-free clauses as the former avoids the problems of data segmentation and it further takes into account every error unlike error-free clauses. These findings lent support to Mehnert (1998), who argued that count-based measures of errors (e.g., *number of errors per 100 words*) were more reliable than percentages of free-errors clauses or units. However, more meta-analyses and comparative studies are still needed to test these assumptions.

As for the measurements of fluency, the six measures failed to show the effects of IR on performance with regards to filled pauses, silent pauses, and repair fluency. While the total number of repairs and filled pauses were higher in the +IR condition, the variation was not significant. The mixed results obtained from the four measures of silent pauses suggested that the impact of manipulating IR on fluency was not captured by these measures. However, these standard measures of fluency are repeatedly used in different studies and are reported as not only reliable indicators of fluency, but they are sensitive to manipulating task design (Tavakoli, 2016). Therefore, the non-significant results obtained in this study proposed that IR would not affect fluency. The similar number and length of silent pauses across the two tasks suggested that the participants’ patterns of silent pauses were the same regardless of the increased

cognitive demands of the tasks. For future studies, the use of pruned and unpruned speech rate measures along with separate measures of repair fluency episodes might shed more light on other facets of fluency. There is research evidence to suggest measures of speech rate are reliable indicators of fluency, and they correlate with language proficiency higher than any other measures of fluency (Tavakoli & Skehan, 2005). Using unpruned speech rate helps to explore the extent to which learners can fill time with speech, whereas pruned speech rate can measure the amount of meaningful speech during a certain period of time. Therefore, combining the two measures will offer valuable data about the quantity and quality of speech produced during specific time under the two conditions, i.e., -IR vs +IR.

5.9 Study One closing remarks

Study One was designed to examine the predictions of Cognition Hypothesis (CH) with respect to the effects of Task Complexity (TC) manipulated by intentional reasoning (IR) on L2 speech performance and perceptions of Task Difficulty (TD). Following the assumptions of the CH, the study hypothesised that TC will push L2 learners to enhance their performance in terms of complexity and accuracy at the cost of fluency. Manipulating TC through IR demands, the current study further aimed to offer a more systematic operationalisation of IR at the levels of task content and instructions. Employing a within-participants design, this small-scale study recruited 20 participants with homogeneous characteristics in terms of age, L1, and LP. Performing two video-based tasks varying in their levels of TC (-/+IR) revealed that the more complex task (+IR) elicited more syntactic complexity and accuracy, but less lexical complexity of oral performance, whereas fluency was not affected by IR requirements. The +IR task was perceived as more difficult than the -IR counterpart.

The results of Study One were discussed with reference to the hypotheses, the previous relevant studies, and measures of analysis. The findings showed mixed support to what was found in previous research regarding the impact of manipulating TC (with focus on IR) on the complexity, accuracy and fluency of L2 performance. However, the findings of TD perceptions were in harmony with previous studies. The findings draw attention to the problematic and inconsistent operationalisations of TC variables such as IR. Thus, there is an urgent need for consensus among L2 researchers on a framework to operationalise and research the variables of TC more systematically and carefully.

The results of the current study evoked some measurement-related issues which need to be considered in future studies. Apparently, the IR demands encouraged the use of more subordinations and longer AS-units, but they did not elicit longer clauses. It is therefore suggested that *mean length of clause* needs to be carefully considered as a measure that taps different aspects of syntactic complexity. Moreover, the +IR demands drove the participants to constantly use more formulaic and repetitive language which affected the type-token ratios in the more complex task. Thus, the low D values in the +IR task did not precisely reflect the complexity of the performance. This is why, it is advised in future studies to use more measures that respond to the different facets of lexical complexity, i.e. size, richness, dispersion and sophistication.

It was also obvious that the extensive use of formulaic units and the repetitiveness which characterised performance in the +IR condition resulted in more accurate speech production which was captured successfully by the two measures of accuracy, i.e. *error-free clauses* and *errors per 100 words*. The variation between the effect sizes of the two measures emphasised the impact of the denominators, i.e. *100 words vs clauses* on the findings. Henceforth, it is recommended combining general and specific measures of accuracy that consider number and percentage of erroneous and correct units. Though, six measures were employed to explore fluency, none of them was adequate enough in capturing any significant variation in terms of repair fluency or breakdown fluency. Accordingly, using measures that reflect speed fluency is highly recommended (e.g., *pruned and unpruned speech rate*) in further research.

Employing a relatively small sample size ($n = 20$) whose participants belong to only one proficiency level (B2) needs to be acknowledged and addressed in future studies. It is recommended that future studies recruit more participants to attain more reliable and generalisable findings and to consider participants with different proficiency levels which offers more in-depth examinations of the other variables that may influence L2 performance. It is possible that the individual difference in learners' variables (e.g. language proficiency) and/or cognitive ability variables (e.g., working memory) contribute to the variations in L2 oral performance on tasks with increased IR requirements. Therefore, a new study is needed to fine-tune a framework to operationalise and investigate IR more carefully and systematically, explore the variables that may moderate the effects of IR, and address the acknowledged limitations of Study One.

5.10 Study Two Proposal

As mentioned earlier, Study One revealed a number of limitations and suggested a wider scope for further research on the effects of IR on L2 performance. Thus, it was inevitable to design a new study that; a) addresses the recommendations of Study One concerning a more thorough investigation of IR and a consideration of the individual differences that may interact with IR; b) considers Study One limitations regarding the choice of the sample size and choice of measurements of analysis.

The complicated picture that has emerged from Study One poses more challenges to gain a better understating of how L2 performance as mediated by variables of TC (e.g. IR demands) promotes language performance. Therefore, a more systematic approach is required to operationalise and investigate IR to depict a comprehensive picture of how task characteristics and conditions interact with learners' individual variables to affect L2 speech production on more complex tasks. Given the limited scope of Study One and the multifaceted nature of TC, it is necessary for Study Two to widen the scope of investigation in order to advance SLA and TBLT research by contributing in offering more robust findings with implications for research and pedagogy.

Individual differences between L2 learners is an area of central importance in task-based and L2 research (Albert, 2011) which still requires further investigations. Given the focus on the effects of IR on L2 performance, it is essential for Study Two to control for these individual variables, explore the role they play, identify how they interact with different task characteristics and conditions, study how they mediate the effect of IR on L2 performance, and examine if they can be considered as reliable predictors of L2 performance. Language proficiency (LP) and working memory (WM) are assumed to play key roles in performing tasks with increased TC and hence affect L2 oral performance and perceptions of TD (Robinson, 2007, Skehan, 2015b). Previous research on TC has explored IR but has not convincingly considered its interaction with language proficiency and working memory. The question how individual differences, i.e. LP and WM, interact with IR demands to affect L2 oral performance has not been answered yet. This study attempts to answer this question and contribute to advance research on TC under the umbrella of SLA and TBLT.

Within the Cognition Hypothesis as a theoretical framework, Study Two can help verify the predictions that complexifying monologic tasks using resource-directing variables will be

associated with joint positive effects on complexity and accuracy and negative effects on fluency. Furthermore, employing participants with different LP can confirm whether the effects of IR on CALF will be the same regardless of learners' LP. It is assumed that the new study will explore whether there are interactions between IR requirements and certain measures of analysis which can be helpful in offering recommendations on what measures are more reliable and sensitive to reflect each aspect of CALF on tasks that require IR, and hence help TBLT research to validate, refine, extend CALF measures.

Task Difficulty will be also investigated using the same TD framework suggested by Study One which includes gathering quantitative and qualitative data from a retrospective questionnaire on the participants' perceptions of TD. The findings of the TD investigation are needed to, 1) broaden knowledge about the distinction between TC and TD; 2) broaden understanding about the factors that contribute to perception of TD; 3) help TBLT research establish an index of TD with further implication for L2 pedagogy and syllabus design; and 4) serve as a tool to validate the effectiveness of IR framework which Study One has put forward.

Finally, Study Two will employ the same framework that Study One has suggested to operationalise intentional reasoning at two levels, 1) *task instructions*, which are assumed to pose task-induced cognitive demands by explicitly requiring the participants to read the characters' intentions, predict their actions and reactions and justify them; and 2) *task content*, which is assumed to pose task-inherent cognitive demands through the unfamiliar, unclear, unpredictable, and reasoning evoking events of the story of the +IR video. Study Two is further designed in an attempt to validate IR framework as a novel contribution of this study which is expected to serve and guide future research on TC.

The following chapters present the methodology and results chapters of Study Two. The two chapters will be followed by a discussion chapter which will discuss the results of the two studies, in addition to other issues that have emerged with regards to the frameworks of TC, models of speech productions, and choice of CALF measurements. A conclusion chapter will then draw conclusions from the findings of Study One and Study Two. The chapter then highlights the original contributions of this thesis and its potential implications for research and pedagogy. Finally, the limitations of this study will be acknowledged and suggestions for future research will be made.

Chapter 6: METHODOLOGY: STUDY TWO

6.1 Introduction

This chapter begins with describing the aims and the research questions of Study Two. Then, the methodological procedures which have been followed to collect, transcribe, code and analyse the data are explained and justified. This includes presenting the study design and the variables under investigation. Then, the participants, tasks, and instruments of the study are introduced. The ethical procedure is then acknowledged, before describing the procedure of data collection in detail. Finally, the chapter describes the process of transcribing, coding and analysing the data.

6.2 Aims of study

Motivated by the mixed findings and recommendations of Study One, a new study was designed with a wider scope and more fine-tuned study design. Study One aimed to examine the impact of manipulating Task Complexity (TC) along Intentional Reasoning (IR) demands on L2 oral performance and perceptions of Task Difficulty (TD). Following up, Study Two was designed to investigate any potential interaction between learners' individual variables and the effects of increasing TC through IR requirements on L2 oral performance due to the assumptions that L2 learners become more dependent on their cognitive and ability individual factors as TC increases, and therefore contribute to language performance and perceptions of TD (Robinson, 2001, 2015).

This line of TC research still suffers paucity of studies that tackle simultaneously the interaction between task variables and learner factors in a more systematic approach. Hence, Study Two is carefully designed to aim at investigating: 1) the effects of manipulating TC by varying degrees of IR in oral narratives on (i) L2 oral performance and (ii) learner perceptions of TD; 2) whether learners' individual differences in (i) Language Proficiency (LP) and (ii) Working Memory (WM) mediate the effects of IR; and 3) to what extent LP and WM can predict performance on tasks requiring different degrees of IR.

6.3 Research Questions (RQ) & Hypotheses (H)

RQ1: What is the effect of TC, manipulated by degree of IR required to complete the tasks, on learners' L2 oral performance, measured by syntactic complexity, lexical complexity, accuracy, and fluency?

H1: Following the predictions of CH, increasing TC along (-/+IR) will direct learners' attention to focus on form to meet the increased cognitive demands, thus promoting syntactic complexity, lexical complexity and accuracy at the cost of fluency.

This question will be addressed through analysing and comparing the participants' oral performance on two video-based narrative tasks which require varying degrees of IR.

RQ.2: Do L2 learners perceive the more complex task as more difficult?

H.2: Following Robinson (2007) and Tavakoli (2009a), it is predicted that learners will perceive the more cognitively complex task i.e. +IR as more difficult than the less complex one, i.e. -IR.

This question will be addressed through a retrospective questionnaire on how learners perceive the tasks in terms of their level of difficulty. The qualitative data gathered about the participants' perceptions of TD will be carefully analysed to know what factors contribute to TD.

RQ3: Does variation in LP mediate the effect of TC, manipulated by IR on the oral performance of L2 learners, measured by syntactic complexity, lexical complexity, accuracy, and fluency?

H3: It is expected that learners at various levels of LP will react to TC in different ways. Concluding from the findings of Study One, it is difficult to predict the direction of the effect.

Q.4: Does variation in WM mediate the effect of TC, manipulated by IR on the oral performance of L2 learners, measured by syntactic complexity, lexical complexity, accuracy, and fluency?

H4: Following the assumption that WM directs and divides attention during online L2 processing (Baddeley, 2012), and that attention is expandable to attend to lexical and morpho-syntactic aspects of L2 and revise semantic-pragmatic concepts (Robinson, 2011a), it is predicted that the effects of WM would be more noticeable in the +IR task (Robinson, 2003). Therefore, it can be anticipated that performance of participants with higher WM will be characterized with higher accuracy (Mota, 2003), syntactic complexity, lexical complexity and speed fluency (Gilbert and Munoz, 2010) in the complex task. However, it is difficult to predict whether WM moderation of TC effects will be significant.

RQ.5: To what extent do LP and WM predict learner speech performance on tasks of different degrees of IR?

H5: It is predicted that there will significant correlations between both LP and WM and the measures of lexical complexity, accuracy, and speed fluency (Gilabert and Munoz, 2010; Mota, 2003). Therefore, it is feasible to predict that variation in LP and WM can explain variation in learners' performance in terms of lexical complexity, accuracy, and speed fluency.

6.4 Study design

A repeated measures within-between-participants factorial design was employed to allow investigating the effect of different variables on the participants' oral performance alongside the interactions between those factors. This analysis was needed since the study draws on a factorial design with three independent variables. IR was a within-participants variable and LP and WM were between-participants variables.

Two tasks were performed with two levels of IR, and performance in the -IR task was compared against performance in the +IR task. TC served as a within-participants variable which was operationalised through the two levels of IR demands. LP and WM were employed as between-participants variables, tested prior to the target story-telling task (see section 6.6 below). Based on the results of the LP and WM tests, the participants were divided into four levels of LP, i.e. A2, B1, B2, C1 (CEFR, Council of Europe, 2001) and three levels of WM, i.e. low, medium and high (Lee & Tedder, 2003; Sagarra, 2008). Therefore, in the current design, IR, LP, and WM were the independent variables in this study, while the dependent variables were the aspects of the participants' L2 oral performance in terms of syntactic complexity, lexical complexity, accuracy and fluency and their perceptions of TD.

Following Study One, this study further employed a mixed methods approach that integrated gathering quantitative and qualitative data to triangulate the investigation and offer more breadth and depth understanding of the relationship between the variables under examination (Creswell, 2015). The design was counterbalanced between participants to control any possible effects of order and practice. Study Two research design and the independent and dependent variables under investigation are summarised in Table 9 below.

Table 9. The study design and variables of Study Two

| Study Design | Independent Variables | Dependent Variables |
|---|---|---|
| <p>Mixed-method within-between-participants design</p> <p style="text-align: center;">↓</p> <p>N = 48</p> | <p>1. Task Complexity (TC) -IR / +IR</p> <p>2. Language Proficiency (LP) A2 / B1 / B2 / C1</p> <p>3. Working Memory (WM) Low / Medium / High</p> | <p>1. L2 speech performance</p> <p style="text-align: center;">↓</p> <p>Syntactic complexity Lexical complexity Accuracy Fluency</p> <p>2. TD perceptions</p> |

6.5 Participants

Study Two recruited 48 learners of English from a secondary school in Jordan. The participants were male, aged 16, and spoke Arabic as a first language. Forty-four participants had never lived in an English-speaking country before. The other four students were born in the United States or the United Kingdom but had returned to Jordan by the age of 5-6, and were therefore deemed appropriate for inclusion. The participants were studying at the same school where data were collected in Study One. The participants were in Year 10 at the time of data collection. The students were placed in different classes depending on their proficiency level in English, which was determined based on the results of different internal English language tests and the class continuous assessments. There are three levels of English language classes at the school. Level A class uses textbooks, which correspond to B2 level based on the Common European Framework of Reference (CEFR). Level B class uses B1-level textbooks, while level C uses textbooks that comply with A2-level. The participants volunteered to take part in the study and they were selected randomly from the three levels. Prior to data collection, their LP was measured using the Oxford Placement Test (Allan, 2004) and an elicited imitation task (Wu & Ortega, 2013), and their WM was measured using backward digit span tests (Kormos & Sáfár, 2008; Morra, 1994; Wright, 2010) in L1 and L2, and these tests will be discussed in the following section.

6.6 Tasks and instruments

6.6.1 The video tasks

Study Two used the same video clips that were adopted in Study One to elicit speech production from the participants individually. The video-based tasks were adopted from *Pat & Mat* (Beneš & Jiránek, 1976), an animated show about two friends who deal with challenges and troubles in optimistic, creative and funny ways. As mentioned in Study One, the choice of the two video clips was validated through a careful selection process involving three researchers who watched a number of episodes before agreeing on selecting two clips. The selection process considered De Jong and Vercelloti's (2015) framework to ensure that the two assigned clips were similar with respect to number of characters, number of elements, duration and storyline, but different regarding the amount of IR required. The selection of the video clips was further validated through the first pilot study, the findings of Study One and a new pilot study.

The only change in Study Two was modifying the duration of the video clips. A decision was made to re-edit some scenes and extend each clip to 120 seconds instead of 90 seconds. This decision was made based on the researcher's observations and the participants' comments in Study One. The participants mentioned that there were quick shifts between some scenes which made it challenging for them to cope with the speed of the storyline and that they were forced to skip narrating parts of some scenes to catch up with the next ones. Therefore, the clips were re-edited to allow more time between some scenes. Piloting the edited clips revealed that the pace of the storyline was suitable and the transition between scenes was smooth. For more detail about the content of video clips and the validation process at the level of content and instructions, see Section 3.6.

6.6.2 Task instructions

Study Two adopted the same task instructions that were piloted and used in Study One. Given that IR was operationalised also at the level of instructions, the instructions explicitly asked the participants to only tell and describe the story in the -IR clip. By contrast, in the +IR task, the instructions encouraged the participants not only to narrate the events of the video clip, but also to read the characters' thoughts and intentions to try to predict and explain their actions and reactions. The instructions were written in both English and Arabic (see Appendix 1 & 2).

6.6.3 Language Proficiency Tests

The Oxford Placement Test (OPT) (Allan, 2004) and an Elicited Imitation Task (EIT) (Wu & Ortega, 2013) were used to measure the participants' LP. Employing both tests is justified as a rich means to tap all aspects of the participants' supposed linguistic knowledge. OPT is assumed to measure L2 learners' explicit knowledge, and EIT is supposed to measure their implicit knowledge (R. Ellis, 2009b; Erlam, 2006). The combined scores of the two tests served as a reference to place the participants into four levels that corresponded to CEFR. It is predicted that using two different tests that measure explicit and implicit language knowledge will offer a more accurate assessment of the learners' overall LP (Elder & R. Ellis, 2009).

The OPT is a reliable and valid test that is widely used in SLA and LT research to allocate students to different levels that correlate with CEFR levels (Geranpayeh, 2003). OPT was first designed by Dave Allan in 1985 and then developed commercially by Cambridge ESOL and Oxford University Press. The test was designed as a quick and convenient tool to measure L2 learners' a) grammatical knowledge of the second language, i.e. grammatical form (accuracy) and grammatical meaning (semantic); and b) pragmatic knowledge, i.e. appropriateness, naturalness, and acceptability of language use (Purpura, 2004). OPT has a computer-based version and a paper-and-pen version. For practical reasons, the pen-and-paper version was administered which included 60 multiple choice questions with a maximum score of 60 points (see Appendix 10). Based on OPT scores, the participants were first placed to CEFR levels as follow: 0-17/A1; 18-29/A2, 30-39/B1; 40-47/B2; 48-54/C1; 55-60/C2 (Allan, 2004). Table 10 below summarises the participants' scores and levels based on OPT results only.

Table 10. Participants' scores in Oxford Placement Test

| Oxford Placement Test | N | Min. | Max. | Mean | Std. Dev. | CEFR |
|----------------------------------|----|------|------|------|-----------|-------|
| Participants' scores Max = 60 | 48 | 20 | 50 | 34.1 | 7.00 | A2-C1 |

Since this study was interested in L2 oral performance, an oral EIT was regarded as a reliable measure of global L2 oral proficiency (Elder & R. Ellis, 2009). EIT is assumed to be an accurate measure of implicit language knowledge (or learners' linguistic competence) and is argued to be a robust tool to discriminate between learners across proficiency levels (Erlam, 2006). Completing an EIT, learners are asked to repeat correctly a set of sentences that get gradually

longer. It is predicted that the longer sentences will surpass the learners' short-term memory, and make it more challenging to repeat the longer ones accurately using explicitly-based recall alone. As a result, L2 learners will rely on their implicit knowledge of L2 semantic and syntactic features to rebuild long sentences subconsciously (ibid). Thus, it is predicted that learners with higher LP and more implicit knowledge will be more successful in completing the task, and hence will eliminate the possible impact of short-term memory assistance during the repeating the utterances. These assumptions were supported by previous research which revealed that EIT correlated positively with standardised proficiency tests but not with WM scores (e.g., Okura & Lonsdale, 2012).

Study Two adopted and adapted an EIT which was developed by Wu & Ortega (2013). Ten meaningful and grammatical sentences were selected with increasing number of syllables, i.e. 8-19 syllables. The first sentence consisted of eight syllables, whereas the last sentence contained 19 syllables. The sentences were recorded by a native speaker of English. Students listened to the recorded sentences and were asked to repeat each one immediately and as accurately as possible. Each sentence was given a score ranged from 0-4 points with 40 points as a maximum total score for the whole task. Each participant was given a maximum of four points for a perfect repetition (repeat the whole sentence correctly) and zero for silence or only a single-word repetition. A copy of the EIT and scoring criteria can be seen in Appendix 11. The participants' scores in the EIT are summarised in Table 11 below.

Table 11. Participants' scores in the elicited imitation task

| Elicited Imitation Task | N | Min. | Max. | Mean | Std. Dev. |
|----------------------------------|----|------|------|------|-----------|
| Participants' scores Max = 40 | 48 | 19 | 40 | 28.3 | 4.86 |

Running Pearson correlation coefficient showed a positive strong correlation between the scores of the OPT and EIT, $r = .65, p = .00$. The scores obtained from the two tests were thus judged suitable to combine to represent the overall learner proficiency. The overall LP scores were also used to divide the participants into levels that corresponded to CEFR by matching the scores against the CEFR descriptors. The OPT contributed to 60% of the total scores and the EIT contributed to 40%. The combined participants' scores in OPT and EIT, i.e. 100 are summarised in Table 12 below.

Table 12. Participants' combined language proficiency scores

| Combined OPT & EIT | N | Min. | Max. | Mean | Std. Dev. |
|-----------------------------------|----|------|------|------|-----------|
| Participants' scores Max = 100 | 48 | 41 | 89 | 62.4 | 10.8 |

As mentioned before, the participants were placed into four groups based on the combined OPT and EIT scores to operationalise their overall LP. The four levels are assumed to correspond to the levels of CEFR. Table 13 below shows the participants' overall LP levels based on their general scores in the two tests, i.e. OPT and EIT.

Table 13. The participants' overall language proficiency levels

| LP Levels | Test scores | N |
|-----------|-------------|----|
| C1 | 76-90 | 5 |
| B2 | 61-75 | 19 |
| B1 | 51-60 | 19 |
| A2 | 41-50 | 5 |

6.6.4 Working Memory Tests

Backward-digit span (BDS) tests (Kormos & Sáfár, 2008; Kormos & Trebits, 2011; Morra, 1994; Wright, 2010) were employed to measure the participants' WM capacity. BDS tasks were preferred over forward-digit counterparts because the former tackle complex verbal WM, which is assumed to involve the central executive capacity and phonological loop, i.e. storage and processing (Gathercole, 1999), while forward-digit span tasks are assumed to only measure the phonological short-term memory. Given the complex mechanisms involved in constructing second language speech and the assumed role of storage/processing trade-off (Wright 2010), the BDS tests are often considered as more appropriate tools in SLA research to measure the phonological loop and the complex WM (Kormos & Trebits, 2011). The complex WM which comprises the central executive function, is argued to control attentional resources (Gathercole, 1999) and spatial and intentional reasoning (Engle et al., 1999), and is therefore well suited to the IR-aspect of the tasks being tested here. Moreover, digit span tests are assumed to be language independent because they are primarily non-verbal tasks. That is, they minimise any impact of learners' linguistic proficiency on learners' performance in WM tests (Wright 2010). BDS tasks in this study were in L1 (Arabic) and L2 (English), to cross-check that they were not affected by the participants' language proficiency.

The WM tests were designed by the researcher. Seven sets of increasing numbers in Arabic and English were audio recorded by the researcher at one digit per second. The first set included 3 digits and the last set consisted of 9 digits. The two versions included different numbers that did not follow any patterns. The participants were required to listen to these sets of increased digits and repeat them backwards. The Arabic and English versions were counterbalanced between participants. The participants were given three attempts for each set. Each participant's WM span was determined based on the last set of digits he repeated successfully twice (Wright 2010). That is if a participant failed to repeat two sets out of three of the same span, his WM would be the last set he repeated successfully twice. Copies of the backward-digit span tests can be found in Appendix 12. The participants' results on the two WM tests are summarised in Table 14 below.

Table 14. Participants' scores in the working memory tests

| Participants' scores on WM tests | N | Min. | Max. | Mean | Std. Dev. |
|----------------------------------|----|------|------|------|-----------|
| BDS in L1 (Arabic) | 48 | 4 | 9 | 5.16 | 1.22 |
| BDS in L2 (English) | 48 | 4 | 9 | 5.02 | 1.17 |

Based on the results of the two BDS tests, the participants' WM in L1 and L2 ranged from four to nine-digit span. A strong and positive correlation was found between L1 and L2 BDS scores, $r = .87, p = .00$. Moreover, a medium correlation was found between L1 WM and overall LP, $r = .38, p = .008$, and between L2 WM and overall LP, $r = .43, p = .002$. Considering the strong correlation between L1 and L2 WM tests, the scores of the L1 BDS test were adopted to group the participants into three levels based on their WM span (Lee & Tedder, 2003; Sagarra, 2008) to enable using WM as an independent variable in further statistical analyses (Wright 2010). Table 15 below shows the three levels of the participants' WM.

Table 15. The participants' working memory levels

| WM Levels | Digit Span | N |
|-----------|------------|----|
| Low WM | 3-4 | 16 |
| Medium WM | 5-6 | 25 |
| High WM | 7-9 | 7 |

6.6.5 Task Difficulty Questionnaire

While the primary aim of this study was to investigate how Task Complexity (TC) could affect L2 oral performance and how individual differences in LP and WM could moderate the effects of TC, understanding learners' perceptions of Task Difficulty (TD) was the secondary aim. Study Two employed the same retrospective questionnaire that was administered in Study One to gather quantitative and qualitative data on the learners' perceptions of TD. The questionnaire was further employed as a validation tool of the operationalisation of IR at the levels of instructions and content. The questionnaire consisted of two questions to rate each task as very easy, easy, difficult or very difficult. Each multiple-choice question was followed by an open-ended question to explain why the participants perceived each task at a certain level of difficulty. See section 3.6, for more detail about the TD questionnaire.

6.7 Pilot study

Since the tasks and the TD questionnaire were piloted in Study one, a new pilot study prior to Study Two was designed to trial the new duration of the video clips, i.e., 120 seconds, and to pilot the WM tests and EIT. The pilot study attempted to examine the practicality of the individual sessions of data collection and to estimate how much time each session would take. Since data were collected individually, it involved each participant to read and sign the ethics forms, complete an L2 background questionnaire, a retrospective questionnaire, do two WM tests, two narrative tasks, and an imitation task, not to mention OPT which would be conducted prior to data collection. There were some concerns about the students' availability and willingness to complete all these tasks individually, and the extent to which they might feel bored, tired or overwhelmed.

Ten overseas students from a University in the UK participated in the pilot study. Their first language was Arabic and they had been living in the UK for about 5-9 months at the time of the pilot study. They were 19-39 years old. Their IELTS scores were 4.5-6.0. Data were collected during one-to-one sessions with the researcher. The participants read and signed the consent form, completed an L2 background questionnaire, read the task instructions, performed the two narrative tasks (-IR/+IR) based on 120-second video clips, did two backward-digit span tasks in L1 and L2 to test their working memory, and an elicited imitation task to test their implicit knowledge of English. Finally, they completed a retrospective questionnaire on their perception of task difficulty. After the end of each session, the researcher had a short interview

with each participant to get more in-depth feedback about the whole session and any practical issues that emerged during performing the tasks.

Piloting the tasks revealed no issues regarding the two-minute duration of the video clips. However, the participants in this pilot study pointed out that the task instructions were not helpful enough to absorb all the tasks requirements. Some of them suggested to do a trial or see a real example before the actual performance. Based on this feedback, a demo video clip from *Pat and Mat* was edited (30 seconds) with the voice of a student telling and describing what was happening. This clip was then shown to each participant during data collection prior to the actual performance as an example along with the written task instructions in Arabic and English. The demo video was expected to help the participants visualise what they were required to do. The decision was made not to give the participants a chance to perform a trial, because this might have a practice effect on their actual performance, and hence influence the results. The findings of the pilot study further suggested that all the participants perceived the task that required more IR as more difficult. The participants attributed their perceptions of TD to the requirements of the +IR task to read thoughts, predict actions, and justify decisions. The content of the +IR video clip was also found to be more difficult to narrate for including events that are less logical and less predictable.

The pilot study offered valuable information regarding the expected time needed for each individual session. The estimated time for each session took about 40-45 minutes for doing all the tasks and completing the questionnaires. This brought forward the issue of practicality that emerged from piloting and simulating the procedure of data collection. There were concerns that 45 minutes would be a long time for the participants to spend on doing several tests and cope with the requirements of the tasks. Besides there could be serious consequences of taking the students out of their classes for about one hour, in terms of causing disruption to the school administration and the feelings of fatigue or lack of interest for the students. In order to overcome these issues and address all these concerns, it was decided to split each data collection session into two meetings over two different days for each participant. In the first meeting, the participants would complete the language background questionnaire and do the two WM tests and the EIT. In the second meeting, they would perform the two video narrative tasks and then complete the TD questionnaire. Thus, each session was expected to last for about 20 minutes which was expected to be a more practical and feasible procedure for data collection.

6.8 Ethical procedures

Study Two followed ethical procedures similar to Study One to ensure that the study and data collection procedure adhered to the University's Ethics Guidance. Prior to data gathering, the researcher obtained ethical clearance from the *School Ethics Committee*. During data gathering, the ethical rights of the participants were assured including their privacy, confidentiality and their right to withdraw from the study at any time. The ethics forms consisted of an information sheet about the study and a consent form (see Appendix 3). Each student read and signed the forms. A copy of the signed consent form was kept with the participants and another copy was kept in the researcher's Department Office. All the learners participated in the study voluntarily but they received pens with the university logo as a gift at the end of the second meeting.

6.9 Data collection procedure

Data was collected from the participants at their school during two individual meetings. The data were collected over a period of three weeks in the spring of 2016. On the first day, the researcher briefed the participants about the study and explained the procedure. The students were asked to read the information sheet and the consent form and sign them before participating in the study. Seventy-one students were randomly selected from Year 10. They signed the ethics forms and sat for a pen-and-paper Oxford Placement Test (OPT). The test was designed to measure the participants' knowledge of English and place them based on CEFR. The participants were randomly divided into three groups and two invigilators were assigned for each group. The invigilators explained the exam instructions and the allocated time, which was 45 minutes. The researcher marked the exam papers using the key answers provided with the test kit. Twenty-three students were later disqualified because they did not complete all the test sections, which showed lack of seriousness or willingness to effectively participate in the study. Based on the results of OPT as a first analysis, the levels of LP of the remaining students were in A2-C1 range based on CEFR, before the overall LP checked against the combined OPT and EIT scores.

Forty-eight students participated in the study and completed all the tasks. A quiet room was prepared by the school administration for the researcher to meet the participants individually. The students were called in an alphabetical order to meet the researcher. Every participant met the researcher individually twice. Each session lasted for about twenty minutes. In the first session, the researcher introduced himself, welcomed the participant and chatted with him for

two minutes to help him relax. The researcher then checked again the student's willingness to participate in the experiment and ensured that he had the right to withdraw from the study at any time. The student completed an L2 background questionnaire.

The first meeting was dedicated for measuring the participants' WM and doing the EIT. The tasks were also counterbalanced between participants to minimize the effect of order and practice. The researcher explained the function of conducting the WM test using back-digit span tasks in English and Arabic. Then, the participant read the instructions of each test and did a trial. The digits with increasing number sets (3-9) were recorded and when the participant was ready, the researcher played the recording. Each participant was requested to repeat each set of digits backwards. Each participant was given three attempts for each set of numbers. While listening, the researcher did the scoring on the participant's task sheet. The test would stop if the participant failed to repeat any set of digits twice. The same procedure was repeated with the other WM test. After that, the participant did the EIT. Each participant listened to ten recorded sentences with increasing length and repeated them as accurately as possible. As in the WM tests, the researcher did the scoring on the participant's task sheet while listening to the participant. The maximum score of the EIT was 40 points.

A second one-to-one meeting was arranged with the same participants on another day to perform two video-based narrative tasks with two levels of IR and complete a retrospective questionnaire on how they perceived the tasks in terms of difficulty. The researcher helped the participant to read and understand the instructions of the first narrative task. The instructions were written in Arabic and English. Then the researcher played the trial video clip (30 seconds), and explained to the participants that he was expected to narrate in a similar way and his voice would be recorded. A voice recording software (Audacity, 2012) installed on the researcher's laptop and a high quality microphone were used to record each performance. When the participant indicated that he was ready, the researcher set the recorder and played one of the video clips. Meanwhile, the participant narrated the story based on the given instructions as he was watching it. Each participant was allowed 20 seconds -if needed- to sum up his narration at the end of each clip. The participants performed the two tasks (-IR/+IR) in a counterbalanced order. After performing the two narrative tasks, each participant completed a questionnaire on his perception of TD, in which each he described how difficult each task was (*very easy – easy – difficult – very difficult*) and justified his perceptions through two open-ended questions.

Once the data were transcribed using SoundScriber software (Breck, 1998), the transcriptions were coded for several measures of CALF. Twenty measures were employed to tap into syntactic complexity, accuracy, lexical complexity, speed fluency, breakdown fluency, and repair fluency. Following the first study, the AS-unit (Foster et al., 2000) was employed to segment the transcriptions into units. To ensure the reliability of data transcription and coding, 10% of the data was checked by an expert researcher, in addition to a native-speaker of English to ensure the reliability of accuracy coding. The next sections provide a full account of the measures used in Study Two to operationalise the four aspects of the learners' oral performance. For more information about the measures employed in Study One, see Section 3.10.

6.9.1 Measures of complexity

Study Two employed the same three measures used in Study One to operationalise syntactic complexity. The three measures were *mean length of AS-unit*, *mean length of clauses*, and *ratio of subordination*. The three measures were adopted to tackle the grammatical complexity with respect to length of clauses and units and ratio of subordination (Norris & Ortega, 2009). AS-unit was adopted to calculate the measures of syntactic complexity for its reliability and validity in analysing oral speech data (Foster et al., 2000). Lexical complexity was represented through a measure of lexical sophistication (PLex Lambda) and lexical diversity (D).

D is a measure of corrected type-token ratio that responds to variation in text length (Malvern & B. Richards, 2002). D is calculated by using Voc-D function available in *Coh-Metrix software* (Graesser et al., 2003). Responding to the limitations of Study One which adopted only one measure of lexical complexity, Study Two included a measure that tapped another dimension of lexis, i.e. sophistication. PLEX Lambda is a measure that assesses the occurrence of using less frequent words in texts produced by L2 learners which weighs their knowledge of more sophisticated words and is assumed to be relatively a reliable and effective tool to analyse short texts (Meara & Bell, 2001). The higher PLEX Lambda is, the rarer and more advanced words are likely to be used in a text. It is therefore predicted that PLEX reflects the learner's advanced levels of vocabulary knowledge (Jarvis 2013). PLEX can be generated using *Lognostics Toolbox*, a free software that offers different tools for researching vocabulary including sophistication (Meara & Bell, 2001). PLEX software splits a text into ten-word sets. Then, it computes the number of less frequent words (not in the 1000 frequent word-list) in each set. The program, then generates a distribution of a parameter, i.e. *lambda*, which ranges

between 0-4 to indicate the extent to which a certain text incorporates more sophisticated and advanced words (Meara & Miralpeix, 2016).

6.9.2 Measures of accuracy

Two measures of accuracy were employed in Study Two, i.e. *percentage of error free clause* (EFC) (Foster & Skehan, 1996) and *Weighted clause ratio* (WCR) (Foster & Wigglesworth, 2016). EFC which was also used in the first study, is a general measure of accuracy that is frequently used in task-based research (R. Ellis and Barkhuizen, 2005). Percentage of EFC is calculated by dividing the number clauses that contain no errors by the total number of clauses produced in a text, multiplied by 100. The higher the EFC percentage is, the more accurate a text is anticipated to be. However, EFC is still criticised for being vulnerable because it is affected by clause segmentation (Mehnert, 1998), clause length (Skehan & Foster, 2005), and error gravity (Foster & Wigglesworth, 2016).

Study Two included a new measure of accuracy, i.e. *weighted clause ratio* (Foster and Wigglesworth, 2016) which gives credit to the accurate language produced by learners rather than only discredits the inaccurate language. Foster and Wigglesworth, in an attempt to address the shortcomings of the general and specific measures of accuracy, propose *weighted clause ratio* (WCR) as a fine-tuned measure that assesses accuracy by weighing errors based on their effect and gravity. That is, WCR gives some credit to less serious errors and to the remaining accurate language that is produced in clauses that contain some errors. Foster and Wigglesworth (2016) propose a scoring scheme that grants *one point* for an accurate clause, *.8 point* for any clause that contains errors that do not affect the message at all, *.5 point* for any error that affects the message to some extent, and *.1 point* for serious errors that completely hinder the message. Then, the total scores of all clauses in a text is divided by the total number of clauses.

However, WCR can be criticized for not making a clear distinction between accurate and inaccurate clauses and also for not addressing the developmental sequences of interlanguage errors across verbal or nominal structures, nor for addressing a potential confound with syntactic complexity. Moreover, the process of coding for WCR can be lengthy and potentially difficult to achieve high inter-rater reliability, as there may be some element of subjectivity in how to rate the effect of different degree of errors on message. However, it is hoped that employing WCR will offer a more in-depth analysis about the levels and types of errors L2 learners make at different LP levels during performing tasks with increased TC.

6.9.3 Measures of fluency

A number of measures were used to capture the different aspects of fluency: that is, the features that break the flow of speech, like pausing and repairs, or that keep the flow of speech, like speed (Skehan, 2014). Study Two employed the same measures of fluency that were used in Study One with respect to breakdown and repair fluency. In order to tackle another important aspect of fluency, two measures of speed fluency were included in Study Two, i.e. unpruned speech rate and pruned speech rate. All the measures of fluency were calculated per one minute.

Measures of pruned and unpruned speech rates are considered as reliable and valid tools to capture global fluency in L2 research (Segalowitz, 2010). Speech rate in this study is calculated based on number of words per 60 seconds. *Unpruned speech rate* (UPSR) is the number of words (including repairs) produced in a sample, divided by the time required to produce that sample, multiplied by 60. *Pruned speech rate* (PSR) is the number of words (excluding repairs) produced in a sample, divided by the time required to produce the sample, multiplied by 60 (Lennon, 1990; Yuan & R. Ellis, 2003). PSR is believed to tackle the three facets of fluency, i.e. speed, pausing and repair, endorsing it as a more reliable and valid measure of speed fluency than UPSR. However, PSR can be criticised for neglecting prominent features of real world speech which is often characterised with incidents of repair (Tavakoli & Foster, 2011).

Researchers using measures of speech rate frequently consider the syllables as the basic units of analysis (Kormos & Denes, 2004). However, following Freed (1995), Lennon (1990), and Skehan & Tavakoli (2005), this study employed words per minute to calculate speech rate. Using word-count as a reference for speech rate is widely accepted in LT research, and is assumed to be more feasible and accurate than counting syllables (Witton-Davies, 2014). It is also claimed that it is more likely that the counted syllables when compared to number of syllables generated through computer programs will not exactly match the number of syllables truly produced (Tauroza & Allison, 1990).

Following Study one, this study employed five measures to capture pausing fluency. That is 1) *mean length of mid-clause silent pauses*, 2) *mean length of end-clause silent pauses*, 3) *number of mid-clause silent pauses*, 4) *number of end-clause silent pauses*, and 5) *number of filled pauses*. Following the non-significant results of the measures of pausing in Study One which used 0.25 second as a threshold, it was decided to set the cut-off point for a silent pause at > 0.40 second (Skehan & Tavakoli, 2005; Tavakoli & Foster, 2011). It is assumed that 0.40

second as a threshold is “brief enough to capture very small interruptions to the speech stream, but long enough to make manual coding feasible” (Skehan, 2014, p. 19). Praat which is a computer software for analysing speech (Boersma & Weenink, 2008) was used to capture and analyse the silent pauses produced by the participants in the data.

Repair fluency was tackled by separate measures of number of *repetitions*, *hesitations*, *reformulations*, *replacements*, and *false starts*, in addition to a composite measure of total number of all types of *repairs* mentioned above. All measures of repairs are counted per 60 seconds. However, these incidents of repairs are claimed to function differently during speech performance. *Repetition* which is repeating the same word(s) and *hesitation* which is repeating part of a word might be used by L2 learners to buy more time for online planning, which can be similar to the role of filled pauses (Duez, 1985). Therefore, it is anticipated that repetition and hesitation can be related to the Conceptualisation and Reformulation stages of Levelt’s model (Witton-Davies, 2014). On the other hand, *reformulation* which is repeating lexical items with modification, *replacement* which is replacing lexical items, and *false start* which is an abandoned utterance are assumed to occur in the monitor component and during the Articulation stage of L2 speech production model as they involve monitoring, modification and self-correction while speaking (Kormos, 2006). Table 16 below summarises the twenty measures used in Study Two to operationalise the various aspects of the learners’ speech performance.

Table 16. Measures of CALF in Study Two

| | Dimension | Measure | Abbrev. | Definition |
|----|-----------------------------|--------------------------------------|----------------|---|
| 1. | Syntactic Complexity | Mean length of AS-units | MLASU | Number of words divided by total number of AS-units. |
| 2. | | Mean length of clauses | MLC | Number of words divided by the total number of clauses. |
| 3. | | Ratio of subordination | ROS | Number of clauses divided by the total number of AS-units. |
| 4. | Lexical | Lexical diversity (D) | D | Adjusted type token ratio computed through Coh-Metrix software |
| 5. | Complexity | Lexical sophistication (PLex Lambda) | PLex | A parameter computed through Lognostics software that reflects number of less frequent words in a sample. |
| 6. | Accuracy | Error-free clause | EFC | Number of error-free clauses, divided by total number of clauses multiplied by 100. |
| 7. | | Weighted clause ratio | WCR | Weighing clauses based on error gravity in a text. The total scores are divided by the total number of clauses. |

| | | | | |
|-----|----------------------------|---|----------------|---|
| 8. | Speed | Unpruned speech rate | UPSR | Number of words (including repairs) produced in a sample, divided by the time required to produce the sample, multiplied by 60. |
| 9. | Fluency | Pruned speech rate | PSR | Number of words (excluding repairs) produced in a sample, divided by the time required to produce the sample, multiplied by 60. |
| 10. | Pausing Fluency | Number of mid-clause silent pauses | NMCSPP | Number of mid-clause pauses over 0.40 s, divided by total time of speech, multiplied by 60. |
| 11. | | Mean length of mid-clause silent pauses | MLMCSPP | Total length of mid-clause pauses over 0.40 s, divided by number of mid-clause pauses over 0.40 s. |
| 12. | | Number of end-clause silent pauses | NECSPP | Number of end-clause pauses over 0.40 s, divided by total time of speech, multiplied by 60. |
| 13. | | Mean length of end-clause silent pauses | MLECSPP | Total length of end-clause pauses over 0.40 s, divided by number of end-clause pauses over 0.40 s. |
| 14. | | Number of filled pauses per minute | NFP | Number of filled pauses such as <i>eh</i> , <i>ah</i> , <i>mm</i> divided by total time of speech in seconds, multiplied by 60. |
| 15. | Repair Fluency | Number of repetitions | NREP | Number of repetitions, divided by total time of speech in seconds, multiplied by 60. |
| 16. | | Number of reformulations | NREF | Number of reformulations, divided by total time of speech in seconds, multiplied by 60. |
| 17. | | Number of replacements | NRPLC | Number of replacements, divided by total time of speech in seconds, multiplied by 60. |
| 18. | | Number of false starts | NFS | Number of false starts, divided by total time of speech in seconds, multiplied by 60. |
| 19. | | Number of hesitations | NHES | Number of hesitations, divided by total time of speech in seconds, multiplied by 60. |
| 20. | | Total number of repairs | NR | Number of all kinds of repairs, divided by the total time of speech in seconds, multiplied by 60. |

6.9.4 Inter-rate reliability

To test the precision of data transcribing and coding, 10% of the data was checked by an expert to assess the inter-rater reliability. Measures of accuracy were also cross-checked by a native speaker expert. Pearson correlation coefficient revealed high agreement between the researcher and the raters with respect to the measures of complexity (94%), accuracy (89%) and fluency (91%). The high inter-rater reliability achieved confirmed the consistency and robustness of the procedure of data transcription and coding. This allowed the researcher to proceed with data analysis with confidence. For samples of the coded data in Study Two, see Appendix 13.

6.10 Data Analysis

Quantitative data analysis was performed using IBM Statistical Package, SPSS 21.0. Descriptive and advanced statistical tests were computed for each of the twenty measures in the two tasks (-IR vs +IR) to examine the participants' oral performance in terms of syntactic complexity, lexical complexity, accuracy and fluency, and hence answer the research questions and test the hypotheses discussed earlier. After checking all the assumptions, a within-participants Multivariate Analysis of Variation (MANOVA) was run to spot any significant differences between the two tasks. Paired-samples t-tests were then run to locate the significant differences, and thus explore the impact of manipulating TC through IR on L2 learners' speech and perceptions of TD. To know whether variations in learners' LP and WM interacted with the effects of IR, two-way between-groups ANOVAs were run. Finally, multiple regression analyses were employed to find out whether LP and WM were reliable predictors of performance on tasks with increased IR demands. Data analysis also included qualitative thematic analyses of the participants' views on their perceptions of TD.

6.11 Conclusion

This chapter presented the aims and research questions of Study Two. This was followed by describing the methodology which included the study design, the participants, the tasks and the instruments employed in data gathering. A detailed description of the pilot study and the ethical procedure were then introduced. The procedures of collecting, transcribing, coding and analysing the data were also discussed. In the next chapter, the data analyses will be presented in detail to answer the research questions of Study Two.

Chapter 7: RESULTS: STUDY TWO

7.1 Introduction

As discussed in the Methodology Chapter, a mixed-methods 2x2 within-between participants factorial design was used in Study Two to examine the effects of the independent variables on several dimensions of participants' oral performance in a story retelling task, and their perceptions of Task Difficulty (TD). Furthermore, the design enabled investigating the possible interaction between these independent variables on L2 oral performance and to what extent they can predict certain aspects of language performance. The independent variables were Task Complexity (TC), language proficiency (LP), and working memory (WM). TC was operationalised through two levels of intentional reasoning, i.e. -IR and +IR. The participants belonged to four levels of LP, i.e., A2, B1, B2, and C1. Their WM span ranged between 4-9 digits and accordingly, they were grouped into three level of WM, i.e. low, medium, and high. The dependent variables were syntactic complexity, lexical complexity, accuracy and fluency of the participants' speech production and their perceptions of TD. All participants performed the two tasks with different degrees of IR, and the two performances were analysed and compared against each other.

Prior to proceeding with advanced inferential analyses, all required assumptions were checked and no violations of any of these assumptions were detected including normality, homogeneity and linearity. Following Study One, a repeated-measures multivariate analysis of variance (MANOVA) was utilized to identify if there were any statistical significant differences between the two performances. The significant statistical results obtained from the MANOVA allowed paired-samples t-tests to be run to identify the location of the significant differences and answer Research Questions 1 and 2 about the effect of TC on oral performance and perceptions of TD. Where significant results were obtained, Cohen's d effect size (Cohen, 1988) were calculated and interpreted using Plonsky and Oswald's (2014) benchmarks. To answer Research Questions 3 and 4, two-way between-groups ANOVAs were used to identify whether LP and WM moderate the effects of TC on distinct aspects of oral language performance. Standard multiple regression analyses were conducted to address Research Question 5 by investigating the interaction between LP and WM and their power in predicting or explaining the participants' language performance in the two tasks (see Research Questions in section 6.3).

7.2 Preliminary analysis

Descriptive analyses were conducted for all the dependent variables under investigation to describe and summarise the basic characteristics of the data gathered in Study Two. The analyses formed a straightforward description of the participants' oral performance and their perceptions of TD. The means and standard deviations are summarised in Table 17 below.

Table 17. Descriptive statistics for the dependent variables of Study Two

| Dimensions | Measures | - IR | | + IR | |
|-----------------------------|---|-------|-------|-------|-------|
| | | Mean | SD | Mean | SD |
| Syntactic Complexity | Mean length of AS unit | 6.77 | 1.23 | 7.56 | 1.06 |
| | Mean length of clauses | 5.21 | .54 | 5.12 | .433 |
| | Ratio of subordination | 1.29 | .16 | 1.47 | .17 |
| Lexical Complexity | Lexical diversity (D) | 25.04 | 10.27 | 23.14 | 8.83 |
| | Lexical sophistication (PLex) | 1.18 | .34 | .85 | .26 |
| Accuracy | Percentage of error free clauses | 43.25 | 17.04 | 57.72 | 15.99 |
| | Weighted Clause Ratio | .79 | .087 | .85 | .06 |
| Speed Fluency | Unpruned speech rate | 105.2 | 24.88 | 117.6 | 26.92 |
| | Pruned speech rate | 92.10 | 23.49 | 103.4 | 26.87 |
| Pausing Fluency | Number of filled pauses | 14.19 | 6.48 | 17.64 | 10.16 |
| | Number of mid-clause silent pauses | 5.66 | 2.84 | 5.37 | 3.07 |
| | Number of end-clause silent pauses | 10.10 | 2.76 | 10.43 | 3.00 |
| | Mean length of mid-clause silent pauses | .94 | .34 | .85 | .28 |
| | Mean length of end-clause silent pauses | 1.30 | .49 | 1.20 | .50 |
| Repair Fluency | Number of repetitions | 4.25 | 3.08 | 4.99 | 3.39 |
| | Number of reformulations | 2.25 | 1.42 | 2.14 | 1.34 |
| | Number of replacements | .80 | .66 | 1.04 | .73 |
| | Number of false starts | .28 | .44 | .32 | .44 |
| | Number of hesitations | 1.28 | 1.07 | 1.28 | 1.20 |
| | Total number of repairs | 8.92 | 4.05 | 9.85 | 4.93 |
| Task difficulty | Perceptions of TD | 1.85 | .54 | 2.72 | .60 |

(*N* = 48)

The means and standard deviations obtained from the descriptive analyses provided a general view of the impact of IR on L2 oral performance and perceptions of TD. The descriptive results suggested that the +IR task elicited language of more syntactic complexity and accuracy than the -IR task. However, less lexically complex language with respect to diversity and sophistication was observed in the +IR task. Regarding fluency, the descriptive results indicated mixed results. Participants in the +IR task produced higher speech rate, more filled pauses and more repairs. Longer mid-clause and end-clause silent pauses were observed in the -IR task, whereas the two tasks elicited the same number of silent pauses mid or end clause position. With respect to self-ratings of TD, the learners perceived the +IR task as more difficult than the -IR task. To find out whether the observed differences between the aspects of performance in the two tasks were statistically meaningful, a multivariate analysis of variance was run.

7.3 Multivariate analysis of variance

Following the descriptive statistics which revealed variations between learners' performances on the -IR/+IR tasks, inferential statistical analyses were run to find out whether these differences were statistically significant. Running MANOVA helped in assessing the variation between the two performances across multiple dependent variables. Statisticians recommend a MANOVA, rather than a set of ANOVAs, as it is a more robust procedure to identify statistical differences in a data set, and to control for Type 1 error (Field, 2013) .

As justified in Study One, it is not preferable to run the MANOVA for all the 20 dependent variables when the sample size is small (Tabachnick & Fidell, 2013). Therefore, only four measures were selected to represent the four aspects of speech performance. The selected measures were reported in the literature (e.g., Skehan & Foster, 2005; Tavakoli & Skehan, 2005) as discrete factors and trustworthy indicators of L2 performance, and hence they were selected to be included in the MANOVA. The measures are:

- 1) *Mean length of AS unit* (Syntactic complexity)
- 2) *Lexical diversity D* (Lexical complexity)
- 3) *Percentage of error free clauses* (Accuracy)
- 4) *Pruned speech rate* (Fluency)

The MANOVA output revealed a statistically significant difference with a large effect size for the four dependent variables combined (Wilks' Lambda = .291; $F = 26.77$, $p = .000$; $\eta^2 = .709$). When considering each dependent variable separately, the differences were also statistically significant in terms of *syntactic complexity* (Wilks' Lambda = .510; $F = 45.22$, $p = .000$; $\eta^2 = .490$), *lexical complexity* (Wilks' Lambda = .913; $F = 4.46$, $p = .04$; $\eta^2 = .087$), *accuracy* (Wilks' Lambda = .386; $F = 74.67$, $p = .000$; $\eta^2 = .614$), and *fluency* (Wilks' Lambda = .605; $F = 30.71$, $p = .000$; $\eta^2 = .395$). The significant results of the MANOVA permitted running paired-sample t-tests to answer Research Questions 1 and 2 regarding the effects of increasing TC through IR demands on L2 learners' oral performance and perceptions of TD.

7.4 Paired-samples t-test

A set of paired-samples t-tests were performed for the twenty measures of CALF and the perceptions of TD to examine whether the differences in mean scores between the two tasks in terms of CALF measures and perceptions of TD reached statistically significant levels. Where significant results were obtained, effect sizes '*Cohen's d*' were calculated (Cohen, 1988). Cohen's *d* is particularly important as a measure that weighs the strength of the significant results of a study. Table 18 below presents the output of the paired-samples t-tests and relevant effect sizes.

The findings obtained from the paired-samples t-tests offered answers to Research Questions 1 & 2 about how some aspects of L2 oral performance and judgement of TD were affected by performing oral narrative tasks of different degrees of IR. To estimate the magnitude of TC effect on L2 performance, effect sizes (Cohen, 1988) were calculated. In terms of interpreting the output of the effect size *d*, this study would go beyond Cohen's interpretation of 0.2 as a small effect size, 0.5, medium, and 0.8 and above, as a large effect size. Study Two adopted the effect size benchmarks suggested by Plonsky and Oswald (2014) who proposed that "Cohen's benchmarks generally underestimate the effects obtained in L2 research (p. 1)". They argued that effect sizes should be interpreted within each discipline, and suggested a new set of threshold levels, i.e. a) $d = 0.4$, a *small* effect size; b) $d = 0.7$, *medium*; and c) $d = 1.0$, a *large* effect size. Thus, Plonsky and Oswald's interpretations of *d* were adopted in the current study. As shown in Table 18 above, the results of the t-tests yielded many (but not universally consistent) statistically significant differences ($p < 0.05$), with different effect sizes. A detailed overview of these findings is presented in the following sections to answer Research Questions 1 and 2.

Table 18. Paired-samples t-tests and effect sizes (Study Two)

| Dimensions | Measures | - IR | + IR | t-test | Sig. (2-tailed) | Effect size |
|---------------------------------|---|-------------------------|-------------------------|--------------|--------------------|----------------|
| | | Mean (SD) | Mean (SD) | <i>t</i> | <i>p</i> | <i>d</i> |
| Syntactic Complexity | Mean length of AS unit | 6.77 (1.23) | 7.56 (1.06) | -5.21 | .000* | .69 |
| | Mean length of clauses | 5.21 (.54) | 5.12 (.43) | 1.37 | .177 | .18 |
| | Ratio of subordination | 1.29 (.16) | 1.47 (.17) | -6.72 | .000* | 1.09 |
| Lexical Complexity | Lexical diversity (D) | 25.04 (10.27) | 23.14 (8.83) | 2.11 | .040* | .20 |
| | Lexical sophistication (PLex) | 1.18 (.34) | .85 (.26) | 6.17 | .000* | 1.09 |
| Accuracy | Percentage of error free clauses | 43.25 (17.04) | 57.72 (15.99) | -8.64 | .000* | .88 |
| | Weighted Clause Ratio | .79 (.08) | .85 (.06) | -5.76 | .000* | .85 |
| Speed Fluency | Unpruned speech rate | 105.2 (24.88) | 117.6 (26.92) | -5.23 | .000* | .48 |
| | Pruned speech rate | 92.1 (23.49) | 103.4 (26.87) | -5.54 | .000* | .45 |
| Pausing Fluency | Number of filled pauses | 14.19 (6.48) | 17.64 (10.16) | -3.26 | .002* | .40 |
| | Number of mid-clause silent pauses | 5.66 (2.84) | 5.37 (3.07) | .653 | .517 | .10 |
| | Number of end-clause silent pauses | 10.10 (2.76) | 10.43 (3.00) | -.948 | .348 | .11 |
| | Mean length of mid-clause silent pauses | .94 (.34) | .85 (.28) | 1.61 | .114 | .29 |
| | Mean length of end-clause silent pauses | 1.30 (.49) | 1.20 (.50) | 1.77 | .082 | .20 |
| Repair Fluency | Number of repetitions | 4.25 (3.08) | 4.99 (3.39) | -1.94 | .058 | .23 |
| | Number of reformulations | 2.25 (1.42) | 2.14 (1.34) | .476 | .636 | .08 |
| | Number of replacements | .80 (.66) | 1.04 (.73) | -1.60 | .116 | .34 |
| | Number of false starts | .28 (.44) | .32 (.44) | -.538 | .593 | .09 |
| | Number of hesitations | 1.28 (1.07) | 1.28 (1.20) | -.040 | .968 | 00 |
| | Total number of repairs | 8.92 (4.05) | 9.85 (4.93) | -1.62 | .111 | .20 |
| TD | Perceptions of task difficulty | 1.85 (.54) | 2.72 (.60) | -7.43 | .000* | 1.52 |

Note: -IR = no reasoning required, +IR = reasoning required, $df = 47$, * p (2-tailed) < 0.05

7.5 Effects of IR on L2 oral performance

The first research question was: “*What is the effect of TC, manipulated by degree of IR required to complete the tasks, on learners’ L2 oral performance, measured by syntactic complexity, lexical complexity, accuracy, and fluency?*” Research Question 1 will be divided into four sub-questions to summarise the results with respect to syntactic complexity, lexical complexity, accuracy and fluency of learners’ L2 oral performance.

7.5.1 Effects of IR on syntactic complexity

For Research Question 1, it was hypothesised that an increase in syntactic complexity would be associated with performing +IR oral narrative tasks. Three measures were employed to operationalise syntactic complexity: *mean length of AS-unit*, *mean length of clauses*, and *ratio of subordination*. The results showed that the participants in the +IR task produced longer AS-units ($M = 7.56$, $SD = 1.06$) than in the -IR task ($M = 6.77$, $SD = 1.23$). The output of the t-test indicated that the difference was statistically significant ($t = -5.21$, $p = .000$) with a medium effect size ($d = .69$). In terms of *mean length of clauses*, although performances in the -IR task generated longer clauses ($M = 5.21$, $SD = .54$) than the +IR task ($M = 5.12$, $SD = .43$), the difference failed to reach a statistical significant level ($t = 1.37$, $p = .177$). Performance on the third measure of syntactic complexity, i.e. *ratio of subordination*, was in favour of the +IR task ($M = 1.47$, $SD = .17$) compared to ($M = 1.29$, $SD = .16$) in the -IR task, and the difference reached a statistically significant level with a large effect size ($t = -6.72$, $p = .000$, $d = 1.09$). The results confirmed the hypothesis with two of the three measures showing that increasing TC using IR demands had a positive impact on syntactic complexity with respect to *mean length of AS-unit* and *ratio of subordination*, but not for *mean length of clauses*.

7.5.2 Effects of IR on lexical complexity

A measure of *lexical diversity* i.e., D , and a measure of *lexical sophistication* i.e., $PLex\ lambda$ were employed to operationalise lexical complexity. It was also predicted that lexical complexity would be affected positively by the IR demands. However, the results showed that it was the less complex task (-IR) that produced more lexically diverse performance ($M = 25.04$, $SD = 10.27$) than the +IR task ($M = 23.14$, $SD = 8.83$). The difference between the two tasks was statistically significant with a small effect size ($t = 2.11$, $p = .040$, $d = .20$). Furthermore, language performance in the -IR task was also characterised by more *lexical sophistication* ($M = 1.18$, $SD = .34$) than the

+IR task ($M = .85, SD = .26$). Variation in lexical sophistication in the two tasks reached a statistical significant level with a large effect size ($t = 6.17, p = .000, d = 1.09$). It could be concluded that the findings for lexical complexity pointed in the opposite direction of the predictions of the Cognition Hypothesis which anticipated positive gains with respect to lexical complexity when performing tasks with increased TC. Producing less diverse and sophisticated language in the +IR task would refute the hypothesis regarding lexical complexity.

7.5.3 Effects of IR on accuracy

It was hypothesised that more accurate language would be associated with a higher IR demand. Two measures were used to represent accuracy in this study. i.e., *percentage of error free clauses* and *weighted clause ratio*. As predicted, the +IR task generated higher *percentage of error free clauses* ($M = 57.72, SD = 15.99$) compared to the -IR task ($M = 43.25, SD = 17.04$). A statistically significant difference was detected between the two tasks with a medium effect size ($t = -8.64, p = .000, d = .88$). This result resembled the one found for *weighted clause ratio*, as learners in the +IR task achieved higher ratio of weighted clauses ($M = .85, SD = .06$) compared to ($M = .79, SD = .08$) in the -IR task. The difference was also statistically significant with a medium effect size ($t = -5.76, p = .000, d = .85$). These findings supported the hypothesis suggesting a positive influence of IR demands on the accuracy of L2 learners' oral performance.

7.5.4 Effects of IR on fluency

For Research Question 1, it was anticipated that increasing TC through IR would result in less fluent language performance, while language in the -IR condition would be more fluent. Different measures were used to respond to the three facets of fluency i.e., speed, breakdown and repairs. *Unpruned and pruned speech rates* were used to measure speed fluency. *Number of filled pauses, number of mid-clause silent pauses, number of end-clause silent pauses, mean length of mid-clause silent pauses, and mean length of end-clause silent pauses* were employed to represent breakdown fluency. Repair fluency was operationalised through separate measures for *number of repetitions, reformulations, replacements, false starts, hesitations*, and a composite measure of *total number of all repairs* produced in each sample. All measures of fluency were calculated per minute.

The +IR task elicited higher *unpruned speech rate* ($M = 117.6, SD = 26.9$) than the -IR counterpart ($M = 105.2, SD = 24.8$). The difference between the two tasks was statistically significant with a small effect size ($t = -5.23, p = .000, d = .48$). Similar results were achieved in terms of *pruned*

speech rate in favour of the +IR task ($M = 103.4$, $SD = 26.8$) compared to ($M = 92.1$, $SD = 23.4$) in the -IR task. The difference reached a statistically significant level with also a small effect size ($t = 5.54$, $p = .000$, $d = .45$). These findings pointed in the opposite direction of the hypothesis with respect to speed fluency which expected lower speech rates in the +IR task.

Performance in terms of breakdown fluency revealed that the participants produced higher *number of filled pauses* in the +IR task ($M = 17.64$, $SD = 10.16$) than in the -IR task ($M = 14.19$, $SD = 6.48$). The variation between the two performances reached a statistically significant level but with a small effect size ($t = -3.26$, $p = .002$, $d = .40$). As regards the four measures of silent pauses, the output of the t-tests revealed mixed results which failed to reach statistically significant levels. Mixed results were also obtained from all measures of repair fluency. Though the total number of repairs showed that the participants produced more repairs while performing the +IR task ($M = 9.85$, $SD = 4.93$) in comparing to the -IR task ($M = 8.92$, $SD = 4.05$), this difference did not reach a statistically significant level. The results obtained for *number of filled pauses* were the only finding to offer support to what was hypothesised, whereas the findings regarding speed fluency measures pointed in the opposite direction of the hypothesis. However, the overall results for the measures of silent pauses and repairs indicated that manipulating the different degrees of IR had no significant effects on silent pausing and repair fluency.

To outline the results of the t-tests with regards to Research Question 1, the findings lent mixed support to what was hypothesised. The hypothesis was confirmed with respect to syntactic complexity and accuracy, whereas the results of lexical complexity and fluency failed to offer support to the predictions. Only the findings with regards to filled pauses were consistent with the predictions that learners would produce more filled pauses in the +IR task. The results, therefore offered empirical evidence that +IR demands would have a negative effect on lexical complexity and a positive effect on speed fluency which contradicted the relevant hypothesis. Building on these findings, it can be proposed that performing tasks with increased IR demands had a positive significant impact on syntactic complexity, accuracy, and speed fluency, a negative effect on filled pausing and lexical complexity, and no significant effects on the learners' performance with respect to silent pausing or repair fluency.

7.6 Effects of IR on perceptions of Task Difficulty

Research Question 2 asked: “*Do L2 learners perceive the more complex task as more difficult?*” This question examined whether learners' perceptions of TD was influenced by increasing TC

through IR demands. It was anticipated that the +IR task would be rated as more difficult than the -IR one. In order to answer this question, data were collected by administering a retrospective questionnaire which asked the participant to rate each task on a four-point scale i.e., (1 = very easy, 2 = easy, 3 = difficult, 4 = very difficult). The questionnaire included two open-ended questions to collect qualitative data to justify the learners' judgements.

The participants rated the +IR task as more difficult ($M = 2.72, SD = .60$) compared to the -IR task ($M = 1.85, SD = .54$). Running paired-samples t-test revealed that the difference between the perceptions of TD in the two tasks reached a statistically significant level with a large effect size in favour of the +IR task ($t = -7.43, p = .000, d = -1.52$). Thus, the hypothesis which predicted that tasks requiring more IR would be rated as more difficult was confirmed. Figure 10 below summarises the participants' perceptions of TD.

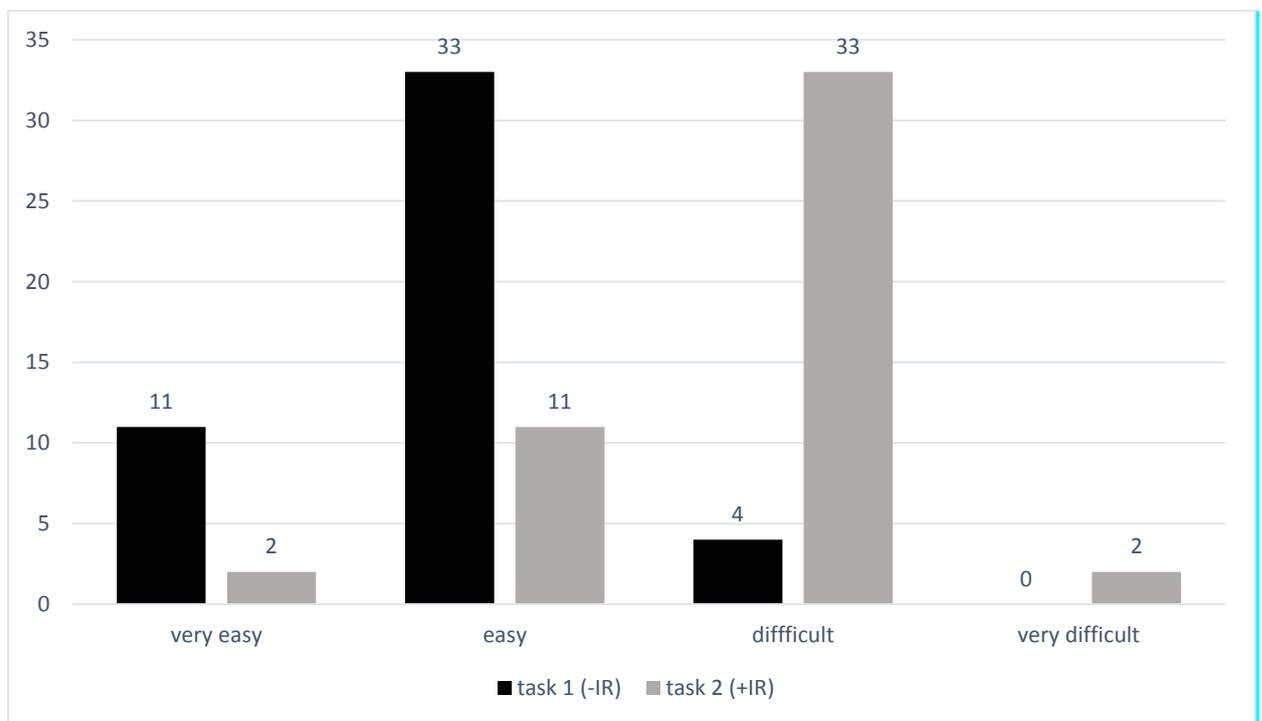


Figure 10. Participants' perceptions of Task Difficulty (Study Two)

Turning to the qualitative data obtained from the two open-ended questions which asked the participants to justify their rating of TD, the thematic analysis revealed four central themes on why the learners perceived the +IR task as more difficult and what task features affected their judgements. The key themes emerging from the participants' answers to the open-ended questions are shown in Table 19 below. The table further shows the frequency and percentages of each theme supported with two examples from the participants' responses.

Table 19. Participants' justifications of TD perceptions (Study Two)

| Themes | Examples from the data | Frequency | % |
|--|---|-----------|------------|
| Cognitive demand (task-induced) caused by task instructions | <i>Difficult because you have to think and predict what is happening and explain their intentions.</i> | 47 | 43% |
| | <i>Easy because describing and telling what's happening is easier than foretelling.</i> | | |
| Cognitive demand (task-inherent, e.g. familiarity, clarity, predictability) | <i>It was difficult because it's not a real-life story and there's a lot imagination in the story.</i> | 45 | 41% |
| | <i>The task was easy because their actions were easy to understand and were predictable.</i> | | |
| Linguistic demands | <i>I find it difficult because I didn't know all the words.</i> | 13 | 12% |
| | <i>Very easy because the words that I used are very easy.</i> | | |
| Time pressure | <i>It is not very easy because sometimes I don't have the time to say the correct explanation of the scene.</i> | 5 | 4% |
| | <i>I needed more time to analyse the events.</i> | | |

Number of responses: 110

Most of the participants attributed their perceptions of difficulty to the greater cognitive requirements imposed in the +IR task. The two main categories that emerged with respect to the increased cognitive demands are *task-induced demands* (triggered by task instructions) and *task-inherent demands* (triggered by task content). High percentage of the responses (43%) denoted task-induced demands to the requirements of the -IR versus +IR tasks as articulated in the instructions of each task. Therefore, the +IR task was perceived as more difficult because it required the participants to predict the characters' behaviours, read their thoughts, explain their intentions, and justify their actions and reactions. On the other hand, many participants found the -IR task easier because they were only required to tell and describe what was happening. The second most frequently mentioned theme was the *task-inherent* cognitive demands (41% of the

comments). The participants regarded their perceptions of difficulty to the degree of topic familiarity, logic, clarity, and predictability in both tasks. For example, a number of participants rated the +IR task as more difficult because the story included unfamiliar, unreal or unpredictable events. Alternatively, those who rated the -IR task as easier, mentioned that the story was easier to predict and narrate because they were familiar with the topic or they found the story plot simpler or clearer or more coherent for containing more logical and real life events than the +IR events.

The linguistic requirements of the -IR and +IR task was the third category that influenced the participants' perceptions of TD receiving 12% of the responses. Some participants mentioned that they needed more complex lexical items or structures to express intentionality or narrate the story of the +IR task. Others mentioned that they could not find the suitable words. Very few comments stated that narrating the -IR task was easier because the required lexical items were easier, simpler, or more frequently used. The final emerging theme was associated with the effects of increasing TC on time pressure. Only 4% of the comments mentioned that time was not enough to meet the cognitive and linguistic demands of the +IR task and that they needed more time to successfully perform the task. Despite the fact that the two tasks were equal in terms of speaking in real time condition, the negative effect of time pressure was more salient in the +IR condition.

Given that the quantitative analysis of the participants' perceptions of TD helped to confirm the hypothesis and answer Research Question 2, the qualitative analysis offered a more in-depth understanding of what aspects of task design contributed to the perceptions of difficulty. The analysis promoted the task cognitive demands imposed by the novel framework of this study which operationalised IR at the levels of task content and instructions as the main contributors to the learners' perceptions of difficulty. However, the emerging themes will be discussed in more detail in the Discussion Chapter conforming to the models of TC and previous research on TD.

7.7 Effects of TC as mediated by language proficiency

Two-way between-groups ANOVAs were conducted to answer Research Question 3: *“Does variation in LP mediate the effect of TC, manipulated by IR on the oral performance of L2 learners, measured by syntactic complexity, lexical complexity, accuracy, and fluency?”* In order to answer this research question, it was necessary to examine the main effect of each independent variable (TC and LP) and their interaction effect on each dependent variable, i.e. CALF measures. It was predicted that the effects of IR condition would be different on learners at different LP levels. However, it was difficult to predict the direction of the effect. The two-way between-groups ANOVAs were

conducted with LP as a between-participants independent variable with four levels (A2, B1, B2, C1), and TC as a within-participants independent variable with two levels (-IR, +IR). The measures that represented syntactic complexity, lexical complexity, accuracy and fluency served as dependent variables. Pallant (2013) argues that “the advantage of using a two-way design is that we can test the *main effect* for each independent variable and also explore the possibility of an *interaction effect*” (p.265). In the current study, the joint effect is likely to occur when the impact of TC on a specific aspect of oral performance depends on the participants’ level of LP.

The ANOVA analyses used Bonferroni adjusted alpha levels of .0125. This adjusted level was obtained by dividing the alpha level (.05) by the number of LP levels (four). Effect sizes of partial eta squared were considered as .01= small, .06 = medium, and .14 = large (Cohen, 1988). Plonsky and Oswald (2014) have attracted attention to the need to carefully interpret measures of the effect sizes based on discipline-specific standards, but since they have not offered interpretations for partial eta squared effect sizes, Cohen’s interpretations will be adopted for the ANOVA meaningful results. Answers to Research Question 3 will be presented in four sub-sections which will summarise the findings of each analysis with respect to the effects of the independent variables individually and jointly on syntactic complexity, lexical complexity, accuracy and fluency.

7.7.1 LP and TC effects on syntactic complexity

Table 20 below displays the descriptive statistics for the measures of syntactic complexity, i.e., *mean length of AS-units*, *mean length of clauses* and *ratio of subordination* with LP as a between-participants independent variable and TC as a within-participants independent variable. The results suggested that A2 proficiency group produced the shortest AS-units in the two tasks, whereas the longest AS-units were produced by B1 group in the -IR task and B2 in the +IR task. Again, A2 group produced the shortest clauses and lowest ratio of subordination in The -IR and +IR tasks. The longest clauses were generated by B2 group in the -IR task and C1 in the +IR task. The highest ratio of subordination was in favour of B1 group in the -IR task and B2 group in the +IR task.

Table 20. Descriptive statistics for syntactic complexity (LP as between-subjects variable)

| Independent variables | | Dependent variables | | | | | | |
|-----------------------|--------|---------------------|------|-------------------|-----|------------------------|-----|----|
| LP | TC | Length of AS-units | | Length of clauses | | Ratio of subordination | | N |
| | | Mean | SD | Mean | SD | Mean | SD | |
| A2 | - IR | 5.79 | .80 | 4.93 | .32 | 1.16 | .12 | 5 |
| | +IR | 6.86 | 1.25 | 4.84 | .56 | 1.41 | .30 | 5 |
| | -/+ IR | 6.33 | 1.14 | 4.88 | .43 | 1.29 | .25 | 10 |
| B1 | - IR | 7.10 | 1.62 | 5.18 | .60 | 1.35 | .19 | 19 |
| | +IR | 7.61 | 1.16 | 5.10 | .47 | 1.48 | .16 | 19 |
| | -/+ IR | 7.35 | 1.41 | 5.14 | .53 | 1.42 | .19 | 38 |
| B2 | - IR | 6.78 | .68 | 5.37 | .51 | 1.26 | .13 | 19 |
| | +IR | 7.73 | .89 | 5.17 | .34 | 1.49 | .18 | 19 |
| | -/+ IR | 7.25 | .92 | 5.27 | .44 | 1.37 | .19 | 38 |
| C1 | - IR | 6.45 | 1.24 | 5.04 | .63 | 1.27 | .14 | 5 |
| | +IR | 7.46 | 1.15 | 5.26 | .49 | 1.41 | .14 | 5 |
| | -/+ IR | 6.96 | 1.25 | 5.15 | .54 | 1.34 | .15 | 10 |
| Total | - IR | 6.77 | 1.23 | 5.21 | .55 | 1.29 | .17 | 48 |
| | +IR | 7.56 | 1.06 | 5.12 | .43 | 1.47 | .18 | 48 |
| | -/+ IR | 7.17 | 1.21 | 5.17 | .49 | 1.38 | .19 | 96 |

To examine the effect of LP and TC and their interaction effect on syntactic complexity of the participants' speech production, two-way between-groups ANOVAs were run. The results are presented in Table 21 below.

Table 21. Effects of LP and TC on syntactic complexity (Two-way ANOVA)

| Source | Length of AS-units | | | Length of clauses | | | Ratio of subordination | | |
|--------|--------------------|--------------|-------------|-------------------|------|----------|------------------------|--------------|-------------|
| | F | Sig. | η^2 | F | Sig. | η^2 | F | Sig. | η^2 |
| LP | 2.328 | .080 | .074 | 1.703 | .172 | .055 | 1.757 | .161 | .057 |
| TC | 9.482 | .003* | .097 | .097 | .757 | .001 | 18.64 | .000* | .175 |
| LP*TC | .320 | .811 | .011 | .471 | .703 | .016 | .727 | .538 | .024 |

* $p < 0.0125$, LP $df(3, 88)$, TC $df(1, 88)$, LP*TC $df(3, 88)$

The outputs of the ANOVAs revealed two statistically significant main effects for TC: one for *mean length of AS-units* with a medium effect size, ($F = 9.482$, $p = .003$, $\eta^2 = .097$), and one for *ratio of subordination* with a large effect size, ($F = 18.64$, $p = .000$, $\eta^2 = .175$). However, no main effect was detected for TC on *mean length of clauses*, ($F = .097$, $p = .757$). The main effect for LP on the

three measures of syntactic complexity failed to reach statistical significant levels. Furthermore, the interaction effect between LP and TC was not statistically significant for *mean length of AS-units*, ($F = .320, p = .811$), *mean length of clauses*, ($F = .471, p = .703$), and *ratio of subordination*, ($F = .727, p = .538$). From these results, it can be concluded that the effects of TC on syntactic complexity were not moderated by the variation in the learners' LP

7.7.2 LP and TC effects on lexical complexity

The descriptive analysis, shown in Table 22, suggested that lexical complexity enhanced steadily across the four proficiency levels with the higher proficiency speakers producing language of higher lexical complexity. However, all proficiency groups produced language of higher lexical complexity in the -IR task compared to the +IR task. A2 group produced the least diverse and sophisticated lexis in both tasks, whereas the highest lexical diversity in the two tasks was established by the highest LP group (C1). However, it was B2 group who produced the most sophisticated words in the -IR task and C1 group in the +IR task.

Table 22. Descriptive statistics for lexical complexity (LP as between-subjects variable)

| Independent variables | | Dependent variables | | | | |
|-----------------------|--------|---------------------|-------|------------------------|-----|--------------|
| LP | TC | Lexical diversity | | Lexical sophistication | | Participants |
| | | Mean | SD | Mean | SD | N |
| A2 | - IR | 13.68 | .714 | 1.13 | .31 | 5 |
| | +IR | 14.17 | 1.44 | .73 | .17 | 5 |
| | -/+ IR | 13.92 | 1.10 | .93 | .32 | 10 |
| B1 | - IR | 22.13 | 10.57 | 1.11 | .37 | 19 |
| | +IR | 20.37 | 8.79 | .85 | .32 | 19 |
| | -/+ IR | 21.25 | 9.63 | .98 | .36 | 38 |
| B2 | - IR | 27.76 | 7.74 | 1.28 | .36 | 19 |
| | +IR | 25.05 | 6.57 | .85 | .24 | 19 |
| | -/+ IR | 26.40 | 7.21 | 1.06 | .37 | 38 |
| C1 | - IR | 37.12 | 6.93 | 1.14 | .14 | 5 |
| | +IR | 35.34 | 5.74 | .94 | .26 | 5 |
| | -/+ IR | 36.23 | 6.07 | 1.03 | .23 | 10 |
| Total | - IR | 25.04 | 10.27 | 1.19 | .34 | 48 |
| | +IR | 23.14 | 8.84 | .85 | .27 | 48 |
| | -/+ IR | 24.09 | 9.58 | 1.02 | .35 | 96 |

Table 23 below shows the results of the two-way ANOVAs which examined the statistical value of the interaction effects of LP and TC on lexical complexity.

Table 23. Effects of LP and TC on lexical complexity (Two-way ANOVA)

| Source | Lexical diversity | | | Lexical sophistication | | |
|--------------|-------------------|--------------|------------------|------------------------|--------------|------------------|
| | F | Sig. | Partial η^2 | F | Sig. | Partial η^2 |
| LP | 15.79 | .000* | .350 | .763 | .518 | .025 |
| TC | .515 | .475 | .006 | 17.23 | .000* | .164 |
| LP*TC | .108 | .955 | .004 | .673 | .571 | .022 |

* $p < 0.0125$, LP $df(3, 88)$, TC $df(1, 88)$, LP*TC $df(3, 88)$

The results indicated a significant effect with a large size for LP on *lexical diversity* ($F = 15.79$, $p = .000$, $\eta^2 = .350$). However, the effect for TC did not reach a statistically significant level ($F = .515$, $p = .475$). The interaction between LP and TC was not statistically significant either ($F = .108$, $p = .995$). As for the effects on *lexical sophistication*, no main effect was found for LP ($F = .763$, $p = .518$). However, the main effect for TC was statistically significant with a large effect size ($F = 17.23$, $p = .000$, $\eta^2 = .164$). No significant interaction effect between LP and TC was established ($F = .673$, $p = .571$). This would suggest, again, that the impact of TC on lexical complexity was not mediated by variation in the learners' LP.

7.7.3 LP and TC effects on accuracy

Means and standard deviations for accuracy measures, summarised in Table 24, indicated that accuracy of performance consistently increased in the higher proficiency levels. For both measures of *percentage of error-free clauses* and *weighted clause ratio* increased from A2 to B1, B2 and C1 levels. The language performance of all proficiency groups was consistently more accurate in the +IR task than the -IR task which was reflected by a higher percentage of error-free clauses and higher weighed clause ratios.

Table 24. Descriptive statistics for accuracy (LP as between-subjects variable)

| Independent variables | | Dependent variables | | | | |
|-----------------------|--------|---------------------|-------|-----------------------|------|--------------|
| LP | TC | Error-free clauses | | Weighted clause ratio | | Participants |
| | | Mean | SD | Mean | SD | N |
| A2 | - IR | 19.79 | 9.92 | .68 | .084 | 5 |
| | +IR | 39.42 | 8.41 | .79 | .054 | 5 |
| | -/+ IR | 29.60 | 13.50 | .74 | .087 | 10 |
| B1 | - IR | 37.63 | 14.67 | .76 | .074 | 19 |
| | +IR | 50.32 | 12.72 | .81 | .058 | 19 |
| | -/+ IR | 43.98 | 14.99 | .79 | .070 | 38 |
| B2 | - IR | 49.43 | 12.75 | .82 | .073 | 19 |
| | +IR | 65.32 | 13.41 | .88 | .058 | 19 |
| | -/+ IR | 57.38 | 15.21 | .85 | .072 | 38 |
| C1 | - IR | 64.54 | 5.45 | .88 | .046 | 5 |
| | +IR | 75.30 | 7.51 | .91 | .032 | 5 |
| | -/+ IR | 69.92 | 8.39 | .89 | .041 | 10 |
| Total | - IR | 43.25 | 17.04 | .79 | .087 | 48 |
| | +IR | 57.72 | 15.99 | .85 | .068 | 48 |
| | -/+ IR | 1.018 | .349 | .82 | .083 | 96 |

By means of two-way ANOVAs, shown in Table 25, statistically significant effects were found for LP and TC on both measures of accuracy. However, even without running this analysis, one can assume that as proficiency increases, accuracy increases as well. So, the main question is not whether proficiency has an impact on accuracy, but the main question is to what extent an increase in LP is reflected by an increase in accuracy. All LP and TC main effects on accuracy were large. LP was found to significantly affect *percentage of error-free clauses*, ($F = 24.24, p = .000, \eta^2 = .452$). TC had also a similar effect ($F = 21.65, p = .000, \eta^2 = .197$). However, the interaction effect was not statistically significant ($F = .315, p = .814$). As for the effects on *weighted clause ratio*, a significant effect was found for LP ($F = 15.33, p = .000, \eta^2 = .343$), and TC ($F = 14.39, p = .000, \eta^2 = .141$). However, the interaction effect between LP and TC was not statistically significant ($F = .679, p = .567$). This would mean that the impact of TC on accuracy was not mediated by variation in the learners' LP.

Table 25. Effects of LP and TC on accuracy (Two-way ANOVA)

| Source | Error-free clauses | | | Weighted clause ratio | | |
|--------|--------------------|-------|------------------|-----------------------|-------|------------------|
| | F | Sig. | Partial η^2 | F | Sig. | Partial η^2 |
| LP | 24.24 | .000* | .452 | 15.33 | .000* | .343 |
| TC | 21.65 | .000* | .197 | 14.39 | .000* | .141 |
| LP*TC | .315 | .814 | .011 | .679 | .567 | .023 |

* $p < 0.0125$, LP $df(3, 88)$, TC $df(1, 88)$, LP*TC $df(3, 88)$

7.7.4 LP and TC effects on fluency

The descriptive analysis for the measures of speed fluency, shown in Table 26, suggested that speech rate enhanced steadily across the four proficiency levels with the higher proficiency speakers producing higher speech rates in the two tasks. For both measures of *pruned speech rate* and *unpruned speech rate* increased steadily from A2 to B1, B2 and C1 levels. The speech rates of all proficiency groups were consistently higher in the +IR task.

Table 26. Descriptive statistics for speed fluency (LP as between-subjects variable)

| Independent variables | | Dependent variables | | | | |
|-----------------------|--------|----------------------|-------|--------------------|-------|----|
| LP | TC | Unpruned speech rate | | Pruned speech rate | | N |
| | | Mean | SD | Mean | SD | |
| A2 | - IR | 79.32 | 24.28 | 65.62 | 19.31 | 5 |
| | +IR | 96.78 | 19.12 | 79.76 | 18.39 | 5 |
| | -/+ IR | 88.05 | 22.56 | 72.69 | 19.28 | 10 |
| B1 | - IR | 103.12 | 22.62 | 87.56 | 20.56 | 19 |
| | +IR | 111.20 | 26.01 | 94.50 | 22.70 | 19 |
| | -/+ IR | 107.16 | 24.39 | 91.03 | 21.65 | 38 |
| B2 | - IR | 109.25 | 24.81 | 98.93 | 23.41 | 19 |
| | +IR | 124.27 | 27.55 | 112.40 | 27.57 | 19 |
| | -/+ IR | 116.76 | 26.96 | 105.66 | 26.13 | 38 |
| C1 | - IR | 124.28 | 14.86 | 109.86 | 12.35 | 5 |
| | +IR | 137.40 | 15.93 | 127.52 | 14.74 | 5 |
| | -/+ IR | 130.84 | 16.09 | 118.69 | 15.84 | 10 |
| Total | - IR | 105.27 | 24.88 | 92.10 | 23.49 | 48 |
| | +IR | 117.60 | 26.92 | 103.49 | 26.87 | 48 |
| | -/+ IR | 111.43 | 26.52 | 97.79 | 25.75 | 96 |

The results of the two-way ANOVAs, summarised in Table 27, revealed a statistical significant effect with a large effect size for LP on *unpruned speech rate* ($F = 6.22, p = .001, \eta^2 = .175$). However, the effects for TC did not reach a statistically significant level ($F = 4.83, p = .030$). The interaction between LP and TC was not significant either ($F = .174, p = .914$). As for the effects on *pruned speech rate*, a significant main effect with a large size was found for LP ($F = 9.66, p = .000, \eta^2 = .248$). No main effect was found for TC ($F = 5.30, p = .024$). The combined effect was also not statistically significant ($F = .227, p = .877$). These results again mean that the effects of TC on speed fluency were not mediated by variations in LP.

Table 27. Effects of LP and TC on speed fluency (Two-way ANOVA)

| Source | Unpruned speech rate | | | Pruned speech rate | | |
|--------------|----------------------|--------------|-------------|--------------------|--------------|-------------|
| | F | Sig. | η^2 | F | Sig. | η^2 |
| LP | 6.22 | .001* | .175 | 9.66 | .000* | .248 |
| TC | 4.83 | .030 | .052 | 5.30 | .024 | .057 |
| LP*TC | .174 | .914 | .006 | .227 | .877 | .008 |

* $p < 0.0125$, LP $df(3, 88)$, TC $df(1, 88)$, LP*TC $df(3, 88)$

Turning to filled pausing and repair fluency, the descriptive analysis, shown in Table 28, indicated that *number of filled pauses* decreased steadily across the four proficiency levels with the higher proficiency speakers producing lower number of filled pauses in the two tasks. All proficiency groups consistently produced higher number of filled pauses in the +IR task. As for repair fluency, only *total number of repairs* was included here, but not the sub measures. B2 group produced the lowest numbers of repairs in both tasks, whereas the highest number was produced by A2 speakers in the +IR task and B1 speakers in the -IR task.

Table 28. Descriptive statistics for filled pauses & repairs

| Independent variables | | Dependent variables | | | | |
|-----------------------|--------|---------------------|-------|------------------|------|----|
| LP | TC | Filled pauses | | Total of repairs | | N |
| | | Mean | SD | Mean | SD | |
| A2 | - IR | 19.61 | 6.03 | 8.73 | 2.89 | 5 |
| | +IR | 28.00 | 13.78 | 13.53 | 2.87 | 5 |
| | -/+ IR | 23.80 | 10.96 | 11.13 | 3.71 | 10 |
| B1 | - IR | 15.17 | 6.84 | 9.41 | 4.72 | 19 |
| | +IR | 18.53 | 10.61 | 10.98 | 5.58 | 19 |
| | -/+ IR | 16.85 | 8.97 | 10.20 | 5.16 | 38 |
| B2 | - IR | 12.35 | 5.84 | 8.38 | 3.97 | 19 |
| | +IR | 15.60 | 7.80 | 8.17 | 4.55 | 19 |
| | -/+ IR | 13.98 | 6.99 | 8.27 | 4.21 | 38 |
| C1 | - IR | 11.99 | 5.40 | 9.39 | 3.35 | 5 |
| | +IR | 11.59 | 6.24 | 8.32 | 2.25 | 5 |
| | -/+ IR | 11.79 | 5.51 | 8.86 | 2.75 | 10 |
| Total | - IR | 14.19 | 6.48 | 8.93 | 4.06 | 48 |
| | +IR | 17.63 | 10.16 | 9.86 | 4.93 | 48 |
| | -/+ IR | 15.91 | 8.65 | 9.39 | 4.51 | 96 |

As shown in Table 29, the two-way ANOVAs revealed that only *number of filled pauses* was affected by LP with a large effect size, ($F = 4.93, p = .003, \eta^2 = .144$), whereas no main effect was found for TC ($F = 3.21, p = .076$). The interaction between LP and TC on their combined effect was not statistically meaningful ($F = .501, p = .683$). As regards *total number of repairs*, no main effect was found for LP ($F = 1.77, p = .159$), or TC ($F = 1.30, p = .258$). The interaction effect was not statistically significant either ($F = 1.07, p = .365$). Based on these findings, it can once more be concluded that the effects of TC on the behaviour of filled pausing and repair fluency were not mediated by the variation in the learners' LP.

Table 29. Effects of LP and TC on filled pauses and repairs (Two-way ANOVA)

| Source | Filled pauses | | | Total of repairs | | |
|--------------|---------------|--------------|-------------|------------------|------|----------|
| | F | Sig. | η^2 | F | Sig. | η^2 |
| LP | 4.93 | .003* | .144 | 1.77 | .159 | .057 |
| TC | 3.21 | .076 | .035 | 1.30 | .258 | .014 |
| LP*TC | .501 | .683 | .017 | 1.07 | .365 | .035 |

* $p < 0.0125$, LP $df (3, 88)$, TC $df (1, 88)$, LP*TC $df (3, 88)$

7.8 Effects of TC as mediated by working memory

Research Question 4 asked, “Does variation in WM mediate the effect of task, manipulated by IR on the oral performance of L2 learners, measured by syntactic complexity, lexical complexity, accuracy, and fluency?” In an attempt to answer this question, two-way between-groups ANOVAs were run with WM as a between-participants independent factor with three levels (low, medium, high) and TC was a within-participants independent factor with two levels (-IR, +IR). The dependent variables were the measures of syntactic complexity, lexical complexity, accuracy and fluency. Bonferroni adjusted alpha levels of .016 was adopted by dividing alpha level (.05) by the three levels of WM. The effect sizes of partial eta squared were adopted and interpreted as .01= small, .06 = medium, and .14 = large (Cohen, 1988).

Running a 2 X 2 between-groups factorial analysis of variance was meant to examine the individual and combined effects of WM and TC on L2 oral performance. The two-way ANOVA allowed to check whether the influence of TC on certain dimensions of oral performance was subject to the participants’ variation in their WM span. To enable running this analysis, WM was transformed into a categorical variable. The decision was made to transform WM into levels rather than linear/ordinal measure because it is most suitable for ANOVA. The participants were placed into three levels based on their scores on the WM backward-digit span test in L1, ranging from 4-9. *Low* WM level comprised participants who scored 4, *medium*, those who scored 5-6, and *high* who scored 7-9. It was predicted that WM would benefit speech performance in term of accuracy, lexical complexity and speed fluency (Gilabert and Munoz, 2010; Mota, 2003). However, no predictions were set for significant joint effects of WM and TC on accuracy, lexical complexity and speed fluency. The following sub-sections will illustrate the results obtained on the effects of the independent factors WM and TC on syntactic complexity, lexical complexity, accuracy and fluency of the participants’ oral performance.

7.8.1 WM and TC effects on syntactic complexity

Table 30 below summarises the descriptive statistics for the measures of syntactic complexity with WM as a between-participants independent factor and TC as a within-participants independent factor. The means and standard deviations revealed that the medium WM group produced the longest AS-units in both tasks. The shortest AS-units were produced by the high WM group in the +IR task, and the low WM group in the -IR task. Again, the medium WM group produced the longest clauses in the -IR and +IR tasks, whereas, the high WM group produced the shortest clauses in the two tasks. As regards subordination, the highest ratio in the -IR task was produced by the high WM group, and in the +IR task by the medium WM group. However, the low WM group produced the lowest ratio of subordination in the two tasks.

Table 30. Descriptive statistics for syntactic complexity (WM as between-subjects)

| Independent variables | | Dependent variables | | | | | | |
|-----------------------|--------|---------------------|------|-------------------|------|------------------------|------|----|
| WM | TC | Length of AS-units | | Length of clauses | | Ratio of subordination | | N |
| | | Mean | SD | Mean | SD | Mean | SD | |
| Low | - IR | 6.60 | .985 | 5.19 | .564 | 1.27 | .139 | 16 |
| | +IR | 7.28 | .918 | 5.11 | .272 | 1.41 | .167 | 16 |
| | -/+ IR | 6.94 | .998 | 5.15 | .437 | 1.34 | .169 | 32 |
| Medium | - IR | 6.95 | 1.39 | 5.34 | .527 | 1.29 | .178 | 25 |
| | +IR | 7.75 | 1.03 | 5.15 | .518 | 1.50 | .184 | 25 |
| | -/+ IR | 7.35 | 1.28 | 5.24 | .526 | 1.40 | .208 | 50 |
| High | - IR | 6.48 | 1.22 | 4.86 | .497 | 1.32 | .190 | 7 |
| | +IR | 7.53 | 1.47 | 5.04 | .446 | 1.48 | .185 | 7 |
| | -/+ IR | 7.01 | 1.41 | 4.95 | .463 | 1.40 | .197 | 14 |
| Total | - IR | 6.77 | 1.23 | 5.22 | .550 | 1.29 | .166 | 48 |
| | +IR | 7.56 | 1.06 | 5.12 | .433 | 1.47 | .179 | 48 |
| | -/+ IR | 7.17 | 1.21 | 5.17 | .495 | 1.38 | .194 | 96 |

As shown in Table 31, two-way ANOVAs were conducted to examine the effects of WM and TC on syntactic complexity and to check for any interaction effects. A statistically significant main effect was observed for TC on *mean length of AS-units* with a medium effect size, ($F = 9.632, p = .003, \eta^2 = .097$), whereas no main effect for WM was detected ($F = 1.367, p = .260$). The interaction effect failed to reach a significant level either ($F = .121, p = .886$). As for *mean length of clause*, no main effect was found for WM ($F = 2.018, p = .139$), or TC ($F = .055, p = .815$), or the combined effect ($F = .767, p = .467$). As regards *ratio of subordination*, no main effect was found for WM ($F = 1.126, p = .329$). However, the main effect for TC achieved a significant level with a large effect size ($F = 17.635, p = .000, \eta^2 = .164$). The joint effect of WM and TC was not statistically significant ($F = .347, p = .708$). These findings led to conclude that the variation in WM capacity did not mediate the effect of TC on syntactic complexity.

Table 31. Effects of WM and TC on syntactic complexity (Two-way ANOVA)

| Source | Length of AS-units | | | Length of clauses | | | Ratio of subordination | | |
|--------------|--------------------|--------------|-------------|-------------------|-------------|-------------|------------------------|--------------|-------------|
| | F | Sig. | η^2 | F | Sig. | η^2 | F | Sig. | η^2 |
| WM | 1.367 | .260 | .029 | 2.018 | .139 | .043 | 1.126 | .329 | .024 |
| TC | 9.632 | .003* | .097 | .055 | .815 | .001 | 17.635 | .000* | .164 |
| WM*TC | .121 | .886 | .003 | .767 | .467 | .017 | .347 | .708 | .008 |

* $p < 0.016$, WM $df(2, 90)$, TC $df(1, 90)$, WM*TC $df(2, 90)$

7.8.2 WM and TC effects on lexical complexity

The descriptive statistics, shown in Table 32, revealed that all WM groups produced language of higher lexical complexity in the -IR task. The high WM group produced the most lexically diverse language in the -IR task, while the medium WM group outperformed the other groups in the +IR task. The least diverse language was generated by the low WM group in both tasks. It was the medium WM group who produced the most sophisticated language performance in the -IR task and the low WM group in the +IR task. However, the low WM group produced the least sophisticated lexis in the -IR task, and in the +IR task it was the high WM group.

Table 32. Descriptive statistics for lexical complexity (WM as between-subjects variable)

| Independent variables | | Dependent variables | | | | |
|-----------------------|--------|---------------------|-------|------------------------|------|--------------|
| WM | TC | Lexical diversity | | Lexical sophistication | | Participants |
| | | Mean | SD | Mean | SD | N |
| Low | - IR | 21.74 | 9.07 | 1.10 | .387 | 16 |
| | +IR | 21.40 | 8.75 | .900 | .310 | 16 |
| | -/+ IR | 21.57 | 8.77 | 1.00 | .360 | 32 |
| Medium | - IR | 25.56 | 10.33 | 1.23 | .328 | 25 |
| | +IR | 24.50 | 8.85 | .842 | .262 | 25 |
| | -/+ IR | 25.03 | 9.54 | 1.03 | .354 | 50 |
| High | - IR | 30.69 | 11.25 | 1.21 | .294 | 7 |
| | +IR | 22.23 | 9.48 | .765 | .177 | 7 |
| | -/+ IR | 26.46 | 10.91 | .990 | .329 | 14 |
| Total | - IR | 25.04 | 10.27 | 1.186 | .343 | 48 |
| | +IR | 23.14 | 8.84 | .850 | .267 | 48 |
| | -/+ IR | 24.09 | 9.58 | 1.018 | .349 | 96 |

As outlined in Table 33, the two-way ANOVAs showed that the effect on *lexical diversity* was not statistically significant for WM ($F = 1.808, p = .170$), TC ($F = 2.195, p = .142$), and their joint effect ($F = .990, p = .376$). As regards *lexical sophistication*, no effect was established for WM ($F = .204, p = .816$), or WM and TC dual effect ($F = 1.186, p = .310$). However, a statistical meaningful effect was only detected for TC on *lexical sophistication* ($F = 23.015, p = .000$) with a large effect size, $\eta^2 = .204$. The obtained findings indicated that WM did not moderate the impact of TC on learners' oral performance in terms of lexical complexity.

Table 33. Effects of WM and TC on lexical complexity (Two-way ANOVA)

| Source | Lexical diversity | | | Lexical sophistication | | |
|--------|-------------------|------|------------------|------------------------|--------------|------------------|
| | F | Sig. | Partial η^2 | F | Sig. | Partial η^2 |
| WM | 1.808 | .170 | .039 | .204 | .816 | .005 |
| TC | 2.195 | .142 | .024 | 23.015 | .000* | .204 |
| WM*TC | .990 | .376 | .022 | 1.186 | .310 | .026 |

* $p < 0.016$, WM $df (2, 90)$, TC $df (1, 90)$, WM*TC $df (2, 90)$

7.8.3 WM and TC effects on accuracy

The descriptive statistics for the accuracy measures, presented in Table 34, indicated that the accuracy of performance improved consistently through the different WM levels. For measure of *weighted clause ratio* increased steadily from low to medium and high WM groups in both tasks. The highest percentage of *error-free clauses* was produced by the high WM group in the -IR task, and by the medium WM group in the +IR task. It was the low WM group which produced the lowest percentage of error-free clauses in the two tasks (-IR/+IR). The language performance of all WM groups was constantly more accurate in the +IR task which was reflected by a higher percentage of *error-free clauses* and higher *weighed clause ratios*.

Table 34. Descriptive statistics for accuracy (WM as between-subjects variable)

| Independent variables | | Dependent variables | | | | Participants N |
|-----------------------|--------|---------------------|--------|-----------------------|------|-------------------|
| WM | TC | Error-free clauses | | Weighted clause ratio | | |
| | | Mean | SD | Mean | SD | |
| Low | - IR | 36.35 | 16.07 | .756 | .082 | 16 |
| | +IR | 49.50 | 14.85 | .815 | .073 | 16 |
| | -/+ IR | 42.92 | 16.62 | .785 | .082 | 32 |
| Medium | - IR | 45.84 | 17.02 | .813 | .086 | 25 |
| | +IR | 62.60 | 14.33 | .865 | .061 | 25 |
| | -/+ IR | 54.22 | 17.73 | .839 | .078 | 50 |
| High | - IR | 49.79 | 16.32 | .818 | .084 | 7 |
| | +IR | 59.10 | 18.64 | .887 | .042 | 7 |
| | -/+ IR | 54.45 | 17.51 | .852 | .073 | 14 |
| Total | - IR | 43.25 | 17.048 | .795 | .087 | 48 |
| | +IR | 57.72 | 15.99 | .852 | .068 | 48 |
| | -/+ IR | 50.49 | 17.97 | .823 | .083 | 96 |

The outputs of the two-way ANOVAs, as shown in Table 35 below, revealed statistical significant individual effects for the independent factors i.e., WM and TC on the measures of accuracy. *Percentage of error-free clauses* was affected significantly with a medium effect size by WM ($F = 5.431, p = .006, \eta^2 = .108$) and TC ($F = 12.404, p = .001, \eta^2 = .121$). However, the interaction effect was not statistically significant ($F = .342, p = .712$). As for the effect on *weighted clause ratio*, a significant and large effect was found for WM ($F = 6.243, p = .003, \eta^2 = .122$), and with a medium effect size for TC ($F = 11.742, p = .001, \eta^2 = .115$). Again, WM and TC interaction effect was not statistically significant ($F = .073, p = .929$). This would mean that the impact of TC on accuracy was not mediated by variation in the learners' WM.

Table 35. Effects of WM and TC on accuracy (Two-way ANOVA)

| Source | Error-free clauses | | | Weighted clause ratio | | |
|--------|--------------------|-------|------------------|-----------------------|-------|------------------|
| | F | Sig. | Partial η^2 | F | Sig. | Partial η^2 |
| WM | 5.431 | .006* | .108 | 6.243 | .003* | .122 |
| TC | 12.404 | .001* | .121 | 11.742 | .001* | .115 |
| WM*TC | .342 | .712 | .008 | .073 | .929 | .002 |

* $p < 0.016$, WM $df (2, 90)$, TC $df (1, 90)$, WM*TC $df (2, 90)$

7.8.4 WM and TC effects on fluency

Table 36 below, illustrates the descriptive statistics for the performance of the three WM groups on the -IR/+IR tasks with respect to speed fluency measured by *pruned speech rate* and *unpruned speech rate*. Participants with medium WM achieved the highest pruned and unpruned speech rates in both tasks. The lowest unpruned speech rate was produced by the high WM group in the -IR task and the low WM group in the +IR task. However, the low WM group produced the lowest pruned speech rate in the two tasks.

Table 36. Descriptive statistics for speed fluency (WM as between-subjects variable)

| Independent variables | | Dependent variables | | | | |
|-----------------------|--------|----------------------|-------|--------------------|-------|----|
| WM | TC | Unpruned speech rate | | Pruned speech rate | | N |
| | | Mean | SD | Mean | SD | |
| Low | - IR | 96.68 | 23.71 | 83.85 | 22.82 | 16 |
| | +IR | 110.14 | 23.92 | 95.08 | 23.64 | 16 |
| | -/+ IR | 103.41 | 24.41 | 89.46 | 23.56 | 32 |
| Medium | - IR | 113.39 | 25.22 | 99.26 | 23.75 | 25 |
| | +IR | 122.35 | 29.54 | 108.13 | 29.66 | 25 |
| | -/+ IR | 117.87 | 27.56 | 103.70 | 26.97 | 50 |
| High | - IR | 95.90 | 18.02 | 85.38 | 17.75 | 7 |
| | +IR | 117.67 | 22.67 | 106.12 | 21.23 | 7 |
| | -/+ IR | 106.78 | 22.69 | 95.75 | 21.66 | 14 |
| Total | - IR | 105.27 | 24.88 | 92.10 | 23.49 | 48 |
| | +IR | 117.60 | 26.92 | 103.49 | 26.87 | 48 |
| | -/+ IR | 111.43 | 26.52 | 97.79 | 25.75 | 96 |

Running two-way ANOVAs, as seen in Table 37, revealed an individual significant impact with a medium effect size only for TC on *unpruned speech rate* ($F = 6.146, p = .015, \eta^2 = .064$). No significant effect was found for WM ($F = 3.426, p = .037$), or the interaction effect ($F = .359, p = .700$). The effect on *pruned speech rate* did not reach significant levels for WM ($F = 3.261, p = .043$), TC ($F = 5.517, p = .021$), or the interaction between the two ($F = .313, p = .732$). This would lead to conclude that speed fluency of L2 oral performance on tasks with increased TC was not moderated by variation in the learners' WM.

Table 37. Effects of WM and TC on speed fluency (Two-way ANOVA)

| Source | Unpruned speech rate | | | Pruned speech rate | | |
|--------------|----------------------|--------------|------------------|--------------------|------|------------------|
| | F | Sig. | Partial η^2 | F | Sig. | Partial η^2 |
| WM | 3.426 | .037 | .071 | 3.261 | .043 | .068 |
| TC | 6.146 | .015* | .064 | 5.517 | .021 | .058 |
| WM*TC | .359 | .700 | .008 | .313 | .732 | .007 |

* $p < 0.016$, WM $df (2, 90)$, TC $df (1, 90)$, WM*TC $df (2, 90)$

As regards *number of filled pauses* and *total number of repairs*, the descriptive statistics in Table 38 showed that as the participants with low and medium WM produced the same *number of filled pauses* in both tasks, the high WM group generated the lowest number of filled pauses in each of the tasks. The medium WM group produced the highest number of repairs in the two tasks, whereas the participants with high WM level produced the least number of repairs.

Table 38. Descriptive statistics for filled pauses and repairs (WM as between-subjects)

| Independent variables | | Dependent variables | | | | |
|-----------------------|--------|---------------------|-------|-------------------------|------|----|
| WM | TC | Filled pauses | | Total number of repairs | | N |
| | | Mean | SD | Mean | SD | |
| Low | - IR | 14.09 | 5.90 | 8.77 | 4.16 | 16 |
| | +IR | 17.78 | 9.74 | 9.77 | 4.19 | 16 |
| | -/+ IR | 15.94 | 8.14 | 9.27 | 4.14 | 32 |
| Medium | - IR | 14.78 | 7.10 | 9.36 | 4.27 | 25 |
| | +IR | 17.74 | 10.80 | 10.39 | 5.62 | 25 |
| | -/+ IR | 16.26 | 9.17 | 9.88 | 4.97 | 50 |
| High | - IR | 12.29 | 5.86 | 7.74 | 3.16 | 7 |
| | +IR | 16.93 | 10.18 | 8.13 | 3.94 | 7 |
| | -/+ IR | 14.61 | 8.33 | 7.93 | 3.43 | 14 |
| Total | - IR | 14.19 | 6.48 | 8.93 | 4.06 | 48 |
| | +IR | 17.63 | 10.16 | 9.86 | 4.93 | 48 |
| | -/+ IR | 15.91 | 8.65 | 9.39 | 4.52 | 96 |

As demonstrated in Table 39, the two-way ANOVAs revealed that the effects on *number of filled pauses* did not reach statistically meaningful levels for WM ($F = .197, p = .822$), TC ($F = 3.44, p = .067$), or their interaction effect ($F = .056, p = .946$). Concerning *total number of repairs*, no main effect was found for WM ($F = 1.015, p = .367$), TC ($F = .574, p = .451$), or the combined effect of WM and TC ($F = .029, p = .946$). Based on these results, it can be concluded that the effects of TC on filled pauses and repairs were not mediated by the variation in the learners' WM.

Table 39. Effects of WM and TC on filled pauses and repairs (Two-way ANOVA)

| Source | Filled pauses | | | No of repairs | | |
|--------|---------------|------|------------------|---------------|------|------------------|
| | F | Sig. | Partial η^2 | F | Sig. | Partial η^2 |
| WM | .197 | .822 | .004 | 1.015 | .367 | .022 |
| TC | 3.440 | .067 | .037 | .574 | .451 | .006 |
| WM*TC | .056 | .946 | .001 | .029 | .972 | .001 |

* $p < 0.016$, WM $df (2, 90)$, TC $df (1, 90)$, WM*TC $df (2, 90)$

7.9 LP and WM as predictors of L2 oral performance

Research Question 5 asked whether LP and WM were reliable predictors of L2 oral performance on tasks of different degrees of TC. It was predicted that LP and WM can predict performance regarding lexical complexity, accuracy and speed fluency. To address this question, multiple regression analyses were performed. The regression analysis is a statistical technique that detects the amount of variance in each dependent variable which can be significantly regarded to the impact of the independent variables (Pallant, 2013). In the regression analysis of this study, LP and WM served as predictors. LP and WM were employed in the regression analysis as continuous variables which is more appropriate to the multiple regression. WM scores which included the composite scores of L1 and L2 WM ranged between 8 and 18 ($M = 10.19, SD = 2.32$). LP scores which included the composite scores of OPT and EIT ranged between 41 and 89 ($M = 62.25, SD = 10.88$). Regarding the dependent variables, composite measures of syntactic complexity, lexical complexity, accuracy, speed fluency and pausing fluency were employed. The composite measure of syntactic complexity comprised *mean length of AS unit*, *mean length of clauses* and *ratio of subordination*. The composite measure of lexical complexity incorporated *D*, a measure of lexical diversity, and *Plex lambda*, a measure of lexical sophistication. *Percentage of error-free clauses* and *weighted clause ratio* formed the composite measure of accuracy. Speed fluency included *pruned* and *unpruned speech rates*. The composite measure of pausing fluency contained *number of filled pauses*, *mid-clause silent pauses*, and *end-clause silent pauses*.

The multiple regression analyses as summarised below in Table 40, show the correlation between the dependent variables as composite scores (CALF) and the two predictors (LP and WM), the regression models which indicate LP-WM combined contributions in explaining the variance in each DV, and finally the coefficients which reveal the individual contribution of each independent variable, i.e., LP and WM in predicting the participants' speech performance.

Table 40. Multiple regressions for LP and WM predicting oral performance

| Outcomes | Predictors | Correlations | | Regression models | | | Coefficients | | | | | | |
|----------------------|------------|--------------|-----|-------------------|----------|----------|--------------|-----------------------|----------|-----------|---------|----------|----------|
| | | DVs | IVs | <i>r</i> | <i>p</i> | <i>F</i> | <i>p</i> | <i>R</i> ² | <i>B</i> | <i>SE</i> | β | <i>t</i> | <i>p</i> |
| Syntactic complexity | LP | | | .123 | .203 | .755 | .476 | .032 | .65 | .57 | .18 | 1.13 | .262 |
| | WM | | | -.069 | .321 | | | | -.18 | .20 | -.15 | -.90 | .373 |
| Lexical complexity | LP | | | .618 | .000* | 13.92 | .000* | .382 | 13.6 | 2.84 | .62 | 4.78 | .000* |
| | WM | | | .259 | .038* | | | | .017 | 1.01 | .002 | .016 | .987 |
| Accuracy | LP | | | .716 | .000* | 23.95 | .000* | .516 | 26 | 4.30 | .70 | 6.04 | .000* |
| | WM | | | .349 | .008* | | | | .82 | 1.53 | .06 | .53 | .593 |
| Speed fluency | LP | | | .477 | .000* | 6.68 | .003* | .229 | 57.82 | 16.90 | .50 | 3.42 | .001* |
| | WM | | | .168 | .127 | | | | -1.54 | 6.01 | -.04 | -.25 | .789 |
| Pausing fluency | LP | | | -.414 | .002* | 4.65 | .015* | .171 | -7.43 | 2.65 | -.42 | -2.80 | .007* |
| | WM | | | -.162 | .136 | | | | .081 | .94 | .01 | .086 | .932 |

* $p < 0.05$, $df (2, 45)$

The two predictor variables, i.e. LP and WM correlated significantly ($r = .471, p = .002$). However, this only modest correlation suggested that the two predictors explain different things in the outcome, due to little or no collinearity, which is more beneficial for the regression model (Field, 2013). The results of the regression analysis with regards to *syntactic complexity* showed that the regression model was not significant, $F (2, 45) = .755, p = .476$. This suggested that the participants'

performance in terms of syntactic complexity could not be explained by variation in their LP and WM. As for *lexical complexity*, a statistically significant regression model was found, $F(2, 45) = 13.92, p = .000$. The amount of variance explained by the model for lexical complexity was 38%. In terms of the contribution of each IV to the model, LP contributed significantly ($p = .000$), whereas WM failed to contribute significantly to the model ($p = .987$).

The regression model in relation to *accuracy* attained a significant level, $F(2, 45) = 23.95, p = .000$, with the amount of 52% of the variance explained by the regression model. Again, LP contributed significantly to the model ($p = .000$), whereas WM contribution was not significant ($p = .593$). For *speed fluency*, the model predicting the participants' speech rates from the variation in LP and WM was also statistically significant, $F(2, 45) = 6.68, p = .003$, explaining 23% of the variance. Once again, LP contributed significantly to the model ($p = .001$), whereas WM did not contribute significantly ($p = .789$). The final regression model examined the extent to which LP and WM reliably explained L2 oral performance regarding pausing fluency measured by *number of filled and silent pauses*. The model achieved a statistically significant level, $F(2, 45) = 4.65, p = .015$. The amount of variance explained by the model for pausing fluency was only 17% with only a significant contribution for LP to the model ($p = .007$), but not for WM ($p = .932$).

In sum, running the multiple regression analyses allowed to answer Research Question 5. That is, “*To what extent do LP and WM predict performance on tasks of different degrees of complexity?*”. The results of the regression analyses showed that except for syntactic complexity, all the models were statistically significant regarding lexical complexity, accuracy, speed fluency, and pausing fluency. The amount of variance explained by these models ranged from 17% to 52%. As for the individual contributions of each IV, only LP contributed significantly to all models (except syntactic complexity), whereas WM failed to contribute significantly to any model. This led to conclude that the model predicting accuracy was the most successful model by explaining 52% of the variance and that only LP as a predictor was more reliable than WM in explaining the participants' oral performance in terms of lexical complexity, accuracy, speed fluency, and pausing fluency. However, WM was found to a limited extent to explain variation in accuracy and lexical complexity due to its significant correlation with these two aspects.

7.10 Summary of the key findings

To summarise the results presented in this chapter, the five research questions and the related main findings are conveyed below.

1. *What is the effect of TC, manipulated by degree of IR required to complete the tasks, on learners' L2 oral performance, measured by syntactic complexity, lexical complexity, accuracy, and fluency?*

Performing tasks with increased TC manipulated by degree of IR demands has systematic positive effects on L2 learners' oral performance with respect to syntactic complexity, accuracy and speed fluency, whereas negative effects are found on lexical complexity and filled pausing. TC as operationalised by IR has no systematic impact on silent pausing and repair fluency.

2. *Do L2 learners perceive the more complex task as more difficult?*

Yes. Learners performing tasks with increased TC as manipulated by degree of IR demands, perceive the more complex task that requires reasoning as more difficult than the task that requires no reasoning.

3. *Does variation in LP mediate the effect of TC, manipulated by IR on the oral performance of L2 learners, measured by syntactic complexity, lexical complexity, accuracy, and fluency?*

No. Though LP has main effects on lexical diversity, accuracy, speed fluency and filled pausing, the effect of TC as manipulated by IR on L2 oral performance is not mediated by the variation in learners' LP.

4. *Does variation in WM mediate the effect of TC, manipulated by IR on the oral performance of L2 learners, measured by syntactic complexity, lexical complexity, accuracy, and fluency?*

No. Though WM has a main effect on accuracy, the effect of TC as manipulated by IR on L2 oral performance is not significantly mediated by the variation in learners' WM.

5. *To what extent do LP and WM predict performance on tasks of different degrees of complexity?*

LP is a reliable predictor in explaining variations in L2 oral performance on tasks of different degrees of TC with respect to lexical complexity, accuracy, speed fluency, and pausing fluency. WM is not considered as a strong predictor in explaining variations in L2 oral performance, but there are indications that WM can predict lexis and accuracy, but to a limited extent. Syntactic complexity cannot be explained by variation in the participants' LP and WM.

Chapter 8: DISCUSSION: STUDY TWO

8.1 Introduction

Study Two was designed as a follow-up experiment for Study One to confirm its findings with respect to the effects of increasing Task Complexity (TC) through intentional reasoning (IR) requirements at the levels of task content and task instructions on L2 learners' oral performance and perceptions of Task Difficulty (TD). Additionally, the current study aimed to investigate the interaction between the effects of TC and variation in learners' language proficiency (LP) and working memory (WM), and the extent to which these two individual variables can reliably predict the learners' oral performance in terms of syntactic complexity, lexical complexity, accuracy, and fluency. To achieve the aims of the study, five research questions were formed and then answered in the Results Chapter. All the results obtained and presented earlier will be exposed to further discussion in this chapter. The data presented in the Results Chapter of Study Two will be reproduced in forms of figures and charts and included here in the Discussion Chapter for ease of reference and to serve as visual support during discussion.

The Discussion Chapter is dedicated to discussing the main findings of Study Two which offered answers to the research questions and tested the relevant hypotheses. First, the findings in relation to Research Question 1 and 2 will be discussed and linked to the results of Study One and other previous studies. The relationship between IR demands and measurements of CALF will be also discussed, before discussing the findings in association to the two prevailing models of TC, i.e. Robinson's Cognition Hypothesis and Skehan's Limited Attention Capacity and two models of speech production, i.e. Kormos (2006) and Levelt (1989). Then, the findings of Research Questions 3-5 about the interaction between TC and the individual differences in LP and WM will be discussed and linked to previous studies on TC and individual differences.

8.2 IR and L2 oral performance

Research Question 1 sought to examine the effect of TC, manipulated by IR demands, on the learners' L2 oral performance, measured by syntactic complexity, lexical complexity, accuracy, and fluency. Following the Cognition Hypothesis (Robinson, 2007), it was predicted that increasing TC through -/+IR demands would result in positive gains with respect to syntactic complexity, lexical complexity and accuracy, and negative gains in terms of fluency. The findings revealed that performing the +IR task increased syntactic complexity, accuracy, speed

fluency, and filled pauses. The +IR demands resulted in decreased lexical complexity, while no effect was found for IR demands on silent pausing or repair fluency.

These results partially backed the assumptions of the CH with regards to syntactic complexity and accuracy, but not for lexical complexity and fluency. This was confirmed through the participants' speech production in the +IR task which was characterised with improved syntactic complexity in terms of *ratio of subordination* and *mean length of AS-units* and more accurate language in terms of *percentage of error-free clauses* and *weighted clause ratio*. However, the obtained findings with respect to lexical complexity were in the opposite direction to the predictions of the CH as the participants' speech production was significantly less varied and less sophisticated than in the -IR task. Similarly, the results achieved by the measures of fluency contradicted the CH, except for one measure, i.e. *number of filled pauses* as the +IR condition produced more filled pauses than the -IR condition as the CH predicted. However, the higher pruned and unpruned speech rate produced in the +IR task and the non-significant results regarding breakdown and repair fluency measures implicated that fluency was not negatively influenced by IR requirements as the CH assumed.

Similar to Study One, the overall results of Study Two were not fully in harmony with the two studies examining the effects of IR on L2 oral performance in terms of CALF, i.e. Ishikawa (2008) and Robinson (2007). Ishikawa (2008), who employed monologic oral narrative tasks based on picture strips stories, found that IR demands increased syntactic complexity, accuracy and lexical complexity, but decreased fluency. Ishikawa's results fully supported the predictions of the CH regarding the impact of IR on L2 speech production. However, only the results of the measures of syntactic complexity and accuracy in Study Two supported those obtained by Ishikawa. However, the results of lexis and fluency were in different directions expect of one measure of fluency, i.e. *number of filled pauses*. Robinson (2007), who used dialogic tasks based on hypothetical situations, found that IR requirements had positive effects only on accuracy, whereas syntactic complexity and fluency were unaffected by IR demands. Moreover, performing the +IR task in Robinson's study resulted in lower lexical complexity measured by TTR. Robinson's (2007) findings supported those obtained in Study Two with regards to accuracy, lexis, silent pausing and repair fluency, but not syntactic complexity, filled pausing, and speed fluency. The lack of consistency between the two studies reported here and the previous studies of Ishikawa and Robinson in their findings could be partially attributed to the unsystematic conceptualisation and operationalisation of IR as a variable to manipulate

cognitive task complexity within the context of TBLT. Given the mixed findings of this study, it is clear that future research is still needed on IR, but the operationalisation and methodological approach taken here is presented as an original contribution, as a framework that could be used again to further investigate IR more thoroughly and systematically.

Comparing the findings of Study Two to previous research manipulating TC using different variables also showed inconsistent patterns in terms of CALF measures. Malicka (2014), who manipulated TC in three monologic oral tasks through number of elements and reasoning demands, found that the more complex tasks generated enhanced accuracy and lexical complexity but reduced fluency, whereas no impact was found on syntactic complexity. Except for accuracy and filled pausing, all Malicka's findings opposed those achieved in Study Two. Albert (2011), who increased TC through the requirements of inventing a story vs only telling a picture story, found that the more complex task produced more accuracy at the expense of lexical complexity and fluency. However, no effect was detected on syntactic complexity. Again, these results resembled those found in Study Two regarding accuracy and lexis but not syntactic complexity and fluency. Levkina and Gilabert (2012) who increased TC through (+) elements and (-) pre-task planning, discovered that TC had positive effects on lexis, negative effects on fluency, while no impact was seen regarding syntactic complexity or accuracy. All Levkina and Gilabert's results contradicted those obtained in Study Two.

The results of Study Two also mismatched Jackson and Suethanapornkul's (2013) meta-analysis of TC research, which revealed negative but negligible effects of TC on syntactic complexity and fluency, and positive but negligible effect on accuracy and lexical complexity. Only the findings in terms of accuracy were consistent with the results of Study Two. Malicka and Sasayama (2017) in a recent TC meta-analysis, found that increasing TC using reasoning demands had positive impact on accuracy and lexis and very small negative impact on fluency. However, no effect was reported on syntactic complexity. Again, this meta-analysis was incompatible with what this study found except for accuracy.

All the studies and meta-analyses discussed above, including this study, revealed that manipulating TC following the current TC models failed to show consistent results that would lend full support to the theoretical frameworks of these models. However, the inconsistency between the findings of this study and what were attained in previous research could be attributed to the differences in the way TC was conceptualised and operationalised. The findings of Study One and Two should be interpreted and generalised with caution due to the way the

tasks were designed and controlled, and the way cognitive TC was operationalised. Employing monologic oral narrative tasks, and operationalising TC through IR demands at the levels of content and instructions, had a direct impact on the obtained results. Thus, it is recommended to review the different ways the current theoretical models conceptualise and operationalise TC, taking into account other possible variables or factors that may interact with TC to affect learners' L2 performance (e.g., individual difference and choice of measurements). This can help future research to investigate TC factors more thoroughly and systematically, in order to achieve more reliable and generalisable results.

The results of Study Two confirmed those obtained in Study One, and the findings of the two studies were very similar despite some differences between each study, indicating good inter-study reliability. The specific differences were that the participants in Study Two varied in their LP (A2-C1) compared to the participants in Study One who belonged only to B2 level. Furthermore, Study Two recruited more participants (N = 48) and more measures (20 measures) compared to 20 participants and 12 measures in Study One. Despite these differences between the two studies, the identical results achieved could be perceived as a validation to this study novel contributions through re-conceptualization and re-operationalisation of IR. However, the findings of the two studies suggest that the framework proposed here to operationalise IR, i.e. task-induced and task-inherent reasoning demands, is a major methodological contributor in shaping the participants' oral performance as captured through measures of CALF as used here, and I therefore discuss why this may be so in the following section.

IR requirements were found to directly influence the cognitive processes (e.g., attention, noticing, and reasoning) involved in speech performance in the +IR task. IR as defined and operationalised in this study required the participants not only to narrate the story of the +IR task, but further to read the characters' thoughts and understand their desires to draw true conclusions about what they intend to do, why and what consequences follow. These higher demands encouraged the participants to involve higher cognitive processing to achieve the outcomes of the more complex task which resulted in certain linguistic patterns that were more salient in the +IR task. With these requirements in mind, the participants' role went beyond only telling and describing the story to analysing the events and having more active role in the story. As a result, the participants' attentional resources were directed more to focus on form in the +IR task compared to the -IR counterpart.

Trying to read the characters' thoughts and desires, the participants needed to use more mental state verbs in the +IR task (e.g., *think, believe, assume, guess, seem, intend*). Attempting to predict the actions of the +IR video, the participants were pushed to use more modal verbs (e.g., *be going to, will*) and adverbs of uncertainty (e.g., *maybe, perhaps*). Furthermore, the participants played more active role while trying to explain the unintentional consequences of the characters' actions. Therefore, they used more interactive and commentary expressions to describe this higher level of intentionality (e.g., *Why are they doing this, What are they trying to do, How stupid they are, This is insane, This is impossible, Oh, my God*), before they made judgements about the possible consequences of the characters' decisions (e.g., *Their stuff will be stolen, They spoiled the roof of the car, How can their car start without an engine*). Consequently, the participants needed to use more coordination conjunctions (e.g., *so, but, because*) to link between what they were describing and what they were explaining.

These linguistic patterns that were observed with higher frequency under +IR condition were formulaic in nature and occurred repetitively in each performance to meet the reasoning demands. As a result, positive gains were detected in terms of syntactic complexity, accuracy and speed fluency at the expense for lexical complexity. On the other hand, the linguistic patterns observed in the -IR task were not of formulaic nature and were not repetitive which explained why the speech performance in the -IR task was characterised with increased lexical complexity at the expense of syntactic complexity, accuracy and speed fluency. These linguistic patterns were captured differently by CALF measures which were employed in this study as will be discussed in the following section.

8.3 IR and CALF measurements

The results of the study suggest that the choice of CALF measures might have an impact on the findings of the study. It appears that some measures were directly affected by the way the variables were operationalised. For example, the way the two studies operationalised IR encouraged the participants to produce more short clauses of formulaic nature in the +IR task (e.g., *I think, I assume, it could be, it seems*) to respond to the required reasoning demands. These formulaic chunks were repeatedly used to help the participants read the characters' thoughts and intentions while narrating the +IR story. The use of such formulaic sequences had a positive impact on accuracy, speed fluency and the ratio of subordination, but a damaging effect on length of clauses and lexical diversity.

Due to the excessive use of formulaic sequences to meet IR demands, higher ratio of subordination and longer AS-units were found at the expense of clause length as captured by the three measures of syntactic complexity. These results which supported those obtained in Study One, confirmed the importance of considering task essentialness for subordination as a fundamental criterion for choosing the measures of analysis in terms of syntactic complexity (Inoue, 2016). It is also worth mentioning that all the studies and meta-analyses discussed earlier, except Ishikawa (2008), found either negative or no effect of TC on syntactic complexity. This verified that the way researchers operationalised TC had a direct impact on learners' speech production with respect to syntactic complexity.

Lexical diversity as captured by D revealed that the task that required more IR generated less varied speech production. The same results were achieved in Study One which suggested that the +IR task instructions and content directed the participants to use more repetitive formulaic language to make predictions and explain the characters' intentionality. This assumption was further confirmed through a post hoc analysis for lexical frequency using Compleat Lexical Tutor (Cobb & Free, 2015). The analysis as summarised in Table 41 showed that mental state verbs (e.g., *think, assume, seem, want, etc.*) were repeatedly produced 505 times in the +IR task, compared to 183 times in the -IR task. The second category of lexis was the logical conjunctions (*so, but, because*) which were repeated in the +IR task 266 times, compared to 100 times in the -IR task. The modal verbs were also repeated more frequently in +IR task (231 times) compared to 101 times in the -IR task. The fourth group of lexis was the adverbs of uncertainty (e.g., *maybe, perhaps, probably, etc.*), which were repeated 171 times in the +IR task, compared to only 49 times in the -IR task.

Table 41. Lexical analysis of +IR vs -IR tasks (Study Two)

| <i>Word types</i> | <i>Number of times (+IR)</i> | <i>Number of times (-IR)</i> |
|-------------------------------|------------------------------|------------------------------|
| Mental state verbs | 505 | 183 |
| Conjunctions | 266 | 100 |
| Modals verbs | 231 | 101 |
| Adverbs of uncertainty | 171 | 49 |
| Total | 1173 | 433 |

Again, this analysis indicated that IR as operationalised in this study resulted in reducing lexical diversity due to the repetitive nature of the lexis used in the +IR task. This highlighted the importance of considering the lexical requirement of tasks when making decisions on what aspects of lexical complexity to tap and what measures to use. Responding to this need as suggested in Study One, a new measure was included in the analysis of lexis in Study Two to tap into lexical sophistication, i.e. PLex. The results of PLex showed that the participants in the +IR task produced less sophisticated language than in the -IR task. These results which contradicted the CH anticipations meant that the IR requirements led the participants to use more frequent words in the +IR task instead of using more infrequent lexis. The lexical frequency analysis discussed above confirmed this assumption as it showed that the participants repeatedly produced more frequent words in the +IR condition (*e.g., mental state verbs, conjunctions, modal verbs*) to meet the IR demands. It was possible that the participants produced more varied and more sophisticated vocabulary in the -IR condition because they were not bound to explain and justify the characters' action or describe their thoughts and intentions. Therefore, instead of being bound to expressions such as "I think" or "It seems" to express their hypothetical justifications, they used their time more freely to narrate a wider range of events happening in the story, and thus used wider range of lexical items.

Apparently, the IR demands imposed by the study not only affected lexical diversity negatively but also influenced lexical sophistication in the same way. However, the non-significant correlation between the measures of diversity and sophistication confirmed that each measure tackled different aspect of lexical complexity' and therefore each measure was affected by the IR requirements differently. This was in line with Lu (2012) who reported that lexical diversity and lexical sophistication are distinct aspects of lexical knowledge. However, it could still be concluded that the negative impact of IR on lexical sophistication was a domino effect of IR demands on lexical diversity.

Moving to accuracy measures, the results as detected by *percentage of error free clauses* and *weighted clause ratio* revealed that speech performance in the +IR task was significantly more accurate than in the -IR task. Again, the results suggested that accuracy was affected in similar ways to syntactic and lexical complexity by the way IR was operationalised in this study. As discussed in Study One, it seemed that through producing more short clauses of a formulaic nature in the +IR condition, the accuracy of the participants' speech performance was significantly higher than that in the -IR condition. It can be proposed that there is an

interrelationship between what the measures of complexity and accuracy capture based on task requirements and essentialness. Hence, any improvement in syntactic complexity due to TC manipulation may be combined with improvement in accuracy as a domino effect. The findings of Study One and Two suggest that as syntactic complexity and accuracy increase, lexical complexity decreases in the +IR task. Still, more empirical studies are needed to confirm the joint effect of using formulaic sequences and repetitiveness of lexis on syntactic complexity, lexical complexity and accuracy as captured through certain CALF measures such as ratio of subordination.

The high correlation between *percentage of error free clauses (EFC)* and *weighted clause ratio (WCR)* was another issue that emerged with respect to accuracy measures. The two measures were found to correlate significantly and highly, $r = .87$, $p = .000$. This high correlation suggested that the two measures tapped into the same aspect of accuracy, even though they rate errors differently. While EFC is assumed to be affected by determining clause boundaries and does not consider error gravity, WCR is not affected by clause boundaries and gives some credit to errors that do not inhibit communication. However, the two measures offered fairly equivalent results, and as such including only one in a study would seem adequate. Given that the WCR is recently developed (Foster & Wigglesworth, 2016), more research evidence is required to evaluate the validity and reliability of what WCR measures, and whether using more structure-based analysis of accuracy (e.g. nominal vs verbal morphosyntax) could be useful.

Turning now to how the measures of fluency interacted with IR demands, only the findings obtained from the measures of speed fluency and filled pauses achieved significant results, but each in a different direction. The participants in the +IR task significantly produced more filled pauses than in the -IR task. These results were the only findings that backed the predictions of the Cognition Hypothesis. Similar findings were achieved in Study One regarding filled pausing but the results failed to reach a significant level. This could suggest an effect of the IR requirements on filled pausing behaviour. It seemed that while the participants were trying to predict the characters' actions and read their intentions and thoughts, they needed to buy more time to deal with the increased cognitive processing demands imposed by +IR demands.

However, the non-significant results of the four measures of silent pauses resembled the findings of Study One. This confirmed that IR demands as operationalised in this study influenced filled pausing behaviour but not silent pausing. It seemed the cognitive demands of justifying others' actions and reading their thoughts pushed the participants to use filled pauses

more consistently than silent pauses to stay connected in speech and buy more time while trying to respond to the increased demands. Another finding obtained in both studies was the non-significant results from the measures of repair fluency. The mixed findings of the separate measures of repairs as well as the composite repair measure confirmed that IR demands had no effect on the participants' patterns of repair. This suggested that task-induced and task-inherent IR demands imposed on the participants in the +IR task were not effective enough to cause them to initiate more repairs than in the -IR task.

Responding to the recommendations of Study one to include measures of speed fluency to the analysis, Study Two employed two measures, i.e. *unpruned* and *pruned speech rate*. The significant results of the two measures indicated that the participants produced higher speech rates in the +IR task which contradicted the assumptions of the CH. It was therefore concluded that the required IR demands combined with the requirement to tell and describe the events in the +IR condition encouraged the participants to fill the time with more speech than in the -IR condition which required only to tell and describe.

Another reason for the higher speech rate in the +IR task could be due to the repetitiveness of lexis used and the nature of the formulaic language produced in the +IR condition. It seemed that using more prefabricated chunks (*I think, I assume, it could be, etc.*) in the +IR task frequently helped the participants attain a higher speech rate during a shorter time. Consequently, this could indicate that the two measures of speech rates were effective in detecting the variation between the two performances with respect to speed fluency.

Discussing the interrelationship between manipulating TC through IR demands and CALF measurement revealed interactions between the language required to explain intentionality and the choice of certain measures. It was evident that the way this study operationalised IR at the levels of instructions and content influenced the linguistic patterns produced in the two tasks, i.e. -IR vs +IR. Consequently, it was apparent that the formulaic nature of the language produced in the two tasks was picked up differently through a number of measures which had a direct impact on the findings. Therefore, it can be concluded that any choice of the analytic measures in terms of CALF needs to consider the predicted linguistic features of the tasks as a possible outcome of TC operationalisation. Still, more empirical studies are needed to confirm any interaction between operationalising TC and the choice of CALF measures to support the findings of this study.

8.4 IR and perceptions of task difficulty

Research Question 2 asked whether the participants would rate the task that required more IR as more difficult than the task that required less. Following the CH, it was anticipated that the +IR task would be perceived as more difficult to perform than the -IR task. A retrospective questionnaire was administered to collect quantitative and qualitative data regarding the participants' perceptions of difficulty. Analysing the quantitative data using paired samples t-tests revealed that the +IR task was rated significantly more difficult with a large effect size ($d = 1.52$). Figure 15 below, compares between the perceptions of TD in the two tasks (-IR versus +IR) in Study Two.

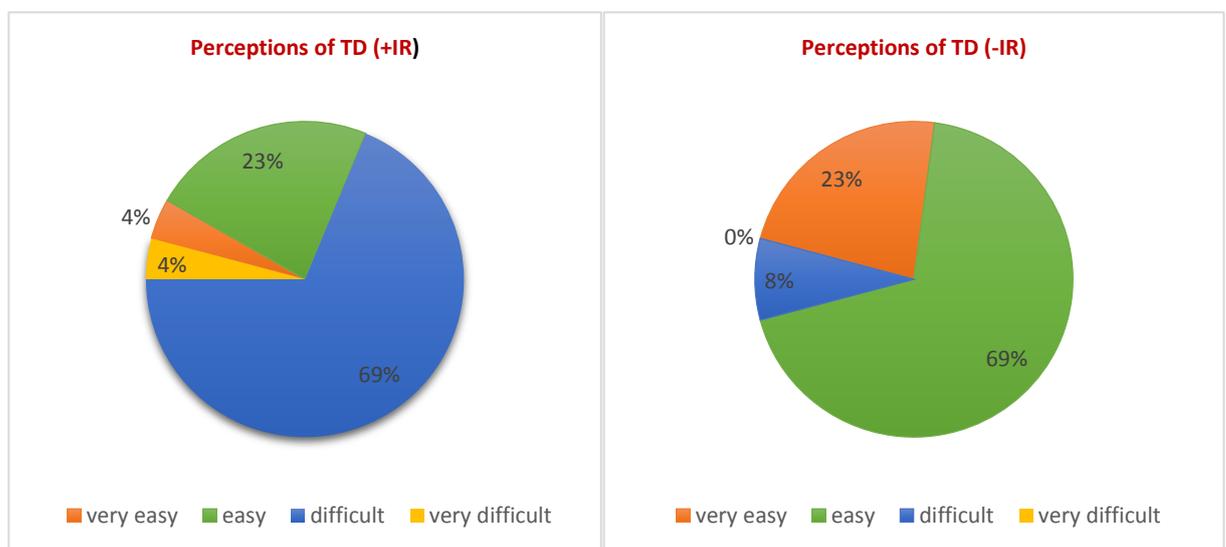


Figure 11. Percentage of Task Difficulty rating (Study Two)

This result lent support to the CH predictions and resembled the findings obtained in Study One. The findings also supported those obtained in previous research (e.g., Gilabert, 2007b; Ishikawa, 2011; Robinson, 2001; Tavakoli, 2009a; Tavakoli & Skehan, 2005; Révész et al., 2016; Robinson, 2007). Following the findings of Study One and Two, it was suggested that perceptions of TD were associated with increased TC through IR demands. The findings obtained from the participants' subjective rating of TD could be further used as evidence to confirm how successful the framework which this study put forward to operationalise IR to manipulate TC. In order to allow a more in-depth investigation of the extent to which TD perceptions could be influenced by this study IR framework that employed intentionality at the levels of task instructions and task content, a qualitative analysis was considered regarding the

participants answers on the two open-ended questions in the questionnaire that encourage the participants to justify their rating of TD.

The qualitative thematic analysis confirmed the findings of Study One. The same two themes i.e. *task-induced* and *task-inherent* cognitive demands were confirmed as the main contributors to the participants' perceptions of TD. Task-induced demands which were encouraged through the task instructions were mentioned as the major factor that affected TD perceptions (43% of the responses). It was evident that asking the participants not only to narrate the story of the +IR video, but also to reason about the characters' intentions and predict their actions affected their perceptions of TD. It was possible that they felt like doing a dual task, i.e. narrating and reasoning which imposed an extra load on their processing and attentional capacities. The second major contributor of increasing the perceptions of TD was task-inherent cognitive demands (41% of the responses) initiated by the variation between the two tasks in their degree of topic familiarity, logic, clarity, and predictability as imposed by the content of each video clip. The participants found that the content of the +IR clip contained less familiar, less logical and less clear events. Furthermore, they found the storyline of the +IR video clip less predictable than the storyline of the -IR video clip. The other two themes but with less contributions to the perceptions of TD were the linguistics requirements (12%) and time pressure (4%). This suggested that failing to retrieve the required lexis or cope with time pressure during oral narrative tasks could be considered as other possible sources for learners' perceptions of TD.

By gaining qualitative information on TD, the findings of this analysis served as a validation to the original contribution of this study in proposing a framework which included TD in the operationalisation of IR at the level of task instructions and task content. The results of Study Two combined with those from Study One for perceptions of TD confirmed the choice of the tasks content and the manipulation of task instructions, and provided useful implications for future research and teaching practices using tasks e.g. in order to avoid inappropriate TD in the classroom (see section 9.4).

8.5 IR and models of task complexity

Research on the effects of manipulating TC on L2 speech production which has underpinned the assumptions in shaping this study, has in general been informed by the predictions of two theoretical models, i.e. Skehan's Limited Attentional Capacity (LAC) and Robinson's

Cognition Hypothesis (CH). As discussed in Chapter 2, LAC assumes that TC consumes the learners' limited attentional resources, resulting in a competition between certain aspects of language performance, i.e. complexity and accuracy. By contrast, CH argues that increasing TC directs the learners' multiple attentional resources to attend to form, i.e. complexity and accuracy at the expense of meaning, i.e. fluency. Hence, LAC speculates a trade-off relationship between accuracy and complexity when performing more complex tasks, while CH advocates that learners can attend to accuracy and complexity simultaneously, benefitting from the availability of multiple pools of non-competing attentional resources.

As noted in section 6.2, this study is testing the effects of manipulating IR on L2 oral performance, focusing on assessing the predictions of Robinson's CH, rather than Skehan's LAC because the former provides a more feasible framework to operationalise TC factors and particularly intentional reasoning. The results of Study One and Two have lent only partial support to the predictions of CH with regards to syntactic complexity and accuracy. However, the findings of the lexical complexity and fluency measures failed to support the hypothesis. Given that lexical complexity pointed in the opposite direction, and breakdown and repair fluency remained unaffected, no decisive conclusions can be reached based on this research on whether complexity and accuracy would advance together or compete against each other when the cognitive demands of a task are increased. Therefore, neither the assumptions of the CH, i.e. joint complexity-accuracy positive gains, nor the assumptions of the LAC, i.e. complexity-accuracy trade-off relationship, are fully confirmed based on the findings of this study.

This conclusion was also supported by Malicka and Sasayama's (2017) meta-analysis which reviewed studies that manipulated TC using resource-directing and resource-dispersing variables. The meta-analysis revealed inconsistent patterns in terms of reasoning demands as a resource-directing variable. Malicka and Sasayama attributed these mixed results to the unsystematic operationalization of reasoning within the CH theoretical framework. Reviewing TC studies that employed the Cognition Hypothesis as a theoretical model, Jackson and Suethanapornkul (2013) also proposed that reasoning demands as a resource-directing variable lacked thorough and systematic operationalisation and conceptualisation that considered the complex nature of reasoning. Further research is clearly still needed to test these predictions more carefully and systematically and to find the reasons for the inconsistency in the findings of studies investigating TC following the current models of TC.

Regarding TD, the predictions of the Cognition Hypothesis that the +IR task will be rated as more difficult is confirmed in this study. Furthermore, these findings support the distinction the CH proposes between TC, which is task-essential, and TD, which is learner-essential. The CH suggests that TC describes within-learner variation in successfully completing any two tasks (e.g., -IR versus +IR tasks), whereas TD, i.e. learners' perceptions of TD, describes between-learner variation in performing the same two tasks (Robinson, 2011a). The latter could be influenced by learners' ability factors (e.g., ability to read others' minds, working memory span, and language proficiency level). This highlights the importance of the individual abilities learners can bring with them to the task which need to be considered, and thereby help researchers establish an index of TD. Although the findings of the tests conducted here were unable to find a strong consistent role for WM and LP in interaction with TC, it is clear that further research is needed to explore the possible interaction between these learners' factors and the aspects of L2 performance on tasks with increased IR requirements. However, this study is an attempt to fill this research gap by being one of the first to systematically investigate the interaction between variations in the learners' language proficiency and working memory and their performance on tasks manipulated by increasing TC through IR demands.

8.6 IR and models of speech production

It is inevitable that any investigation of the effects of TC on L2 oral performance needs to understand and analyse the psycholinguistic processes underpinning L2 speech production (Skehan, 2014) in links with the models of speech production (Kormos, 2006 and Levelt, 1989). Levelt's (1989) definition of a speaker as "a highly complex information processor who can, in some rather mysterious way, transform intentions, thoughts, feelings into fluently articulated speech" (p. 1), can be used as a starting point to discuss how IR demands affect learners' language performance during the different stages of speech production. As mentioned in Chapter 2, Levelt's model of L1 and Kormos's model of L2 production have the same three stages of speech production, i.e. Conceptualisation, Formulation, and Articulation but they differ on how speakers process their speech at each stage. It is assumed that the three stages function simultaneously and effortlessly during L1 speech, but convert to sequential and effortful processing during L2 speech (Skehan, 2014). Based on these stages, complexity is shaped during message planning at the Conceptualisation stage, whereas accuracy and fluency are established during translating the conceptual message into linguistic structures at the

Formulation phase (Levelt, 1989). Introducing IR, as done in a novel way in this study, was intended to test how far and at what point IR might affect the speech production process.

Kormos (2011) argues that the predictions of the CH can be backed by the theoretical frameworks proposed by the models of speech production. Kormos claims that TC activates the learners' L1 and L2 conceptual preparations differently. Within an L2 model of speech production, Kormos (2011) explains that performing more complex tasks manipulated through resource-directing variables (e.g., reasoning) will encourage the learners to stretch their L2 conceptual system to meet the cognitive demands linguistically. As a result, more developed units of lexis, syntax, and morphology will be elicited to express form-meaning associations in L2. The enhancement of the "L2 conceptual system" combined with the activation of "lexical, syntactic, and morphological development" (ibid, p.39) is assumed to result in positive gains in terms of SLA.

From the findings reported here, it seemed that the IR requirements in this study enforced different types of demands on each stage of speech production processing. We suggest that the content-driven IR demands (task-inherent) made the +IR condition more challenging at the planning stage of the Conceptualisation, while the instruction-driven demands (task-induced) taxed the non-verbal and verbal messages at both the Conceptualisation and Formulation stages. It was apparent that justifying others' actions and reasoning about their intentions put pressures on planning and processing language performance and directed the participants to specific linguistic characteristics (e.g. formulaic chunks, short clauses, mental state verbs, logical connectors). Consequently, syntactic complexity, lexical complexity, accuracy and speed fluency were largely influenced by these cognitive demands. Furthermore, it is arguably evident that the greater demands combined across the two stages of speech planning, i.e. Conceptualisation and Formulation had a direct impact on the learners' perceptions of TD as they rated narrating the +IR video clip as more difficult due to task-inherent and task-induced cognitive demands.

The triggering effects of IR demands of L2 concepts and its positive consequences on L2 performance and development offer a good explanation of the significant results with medium and large effect sizes obtained in Study One and Two regarding syntactic complexity, lexical complexity, accuracy, and speed fluency but not breakdown or repair fluency. Despite the reverse direction of lexical complexity results, the post-hoc frequency analysis confirmed that +IR demands triggered the use of more developed lexical items and formulaic chunks (e.g.

verbs of mental states and adverbs of uncertainty). It can be also concluded that due to the need to use more formulaic units to explain form-meaning relations to express intentionality in the +IR condition, the accuracy and speed fluency of the participants improved significantly.

Skehan (2014) stresses the importance of Levelt's model of L1 speech production as a basis to understand why and how TC affects L2 performance. He argues that "in order to hypothesise the effects of task complexity or difficulty on performance, it is essential to understand the kinds of psycholinguistic processes that underline the production of linear stream of speech." (ibid, p.5). Skehan explains that the accuracy of L2 learners drops in more complex tasks because learners are likely to fail in simultaneously attend the preverbal stage (Conceptualisation) and the verbal stage (Formulation). Though Skehan's claims are not confirmed in this study, it is apparent that employing the models of speech production is inevitable to understand the way IR demands affect L2 speech production. Hence, observing the results of this study from a psycholinguistic perspective clearly encourages further investigations about what is going on during each stage of speech processing while performing tasks with increased TC.

8.7 LP and TC interaction effects on L2 performance

Research Question 3 was set to explore whether variation between the participants in LP mediated the effect of variation in TC, manipulated by IR on the oral performance of L2 learners, measured by syntactic complexity, lexical complexity, accuracy, and fluency. It was predicted that at different levels of LP, the participants would respond to TC differently but no hypothesis was set regarding the direction of the effect. To answer this research question, two-way between-groups ANOVAs were conducted with LP as a between-participants variable with four levels (A2, B1, B2, C1), and TC as a within-participants variable with two levels (-IR, +IR). Despite detecting several main effects for LP and TC on the participants' language output, no interaction effects were established between LP and TC on L2 oral performance. This would suggest that performing more complex tasks was not moderated by the variation in the learners' LP. One possible explanation to the non-significant interactions between TC and LP could be regarded to the fact that the participants' performance might be affected by task-inherent and task-induced demands in a way that reduced the moderation of LP. Another possible explanation could be due to the ceiling effect of TC in case of higher proficient learners and the

level beyond which the impact of TC could be spotted in case of lower proficient learners (Kormos & Trebits, 2011).

These findings supported those obtained by Kuiken and Vedder (2008) who found no interaction between LP (low-high) and TC (number of elements) on L2 writing performance, despite finding main effects for LP on complexity and accuracy. Furthermore, Ishikawa (2006) found no interaction effect between LP (low-high) and TC (+/-Here-and-Now) on syntactic complexity and fluency but “there were some indications of interactions between task complexity and proficiency on accuracy and lexical complexity” (p. 212). However, the results of this study did not support those obtained by Malicka and Levkina (2012), who found that LP interacted differently with TC demands as the high-proficiency group in their study produced higher complexity and accuracy, whereas the low-proficiency group produced higher fluency. The non-significant findings with respect to repair fluency in this study were in harmony with Declerck and Kormos (2012) and Gilabert (2007a) who confirmed that repairs behaviour was not affected by variation in LP. It is worth noting that the aforementioned studies operationalised TC differently and employed only two LP groups. The following subsections discuss the findings in more detail with regards to the interactions between the variation in learners’ LP and the effects of manipulating TC using IR demands on each dimension of the L2 oral performance, i.e. syntactic complexity, lexical complexity, accuracy, and fluency. As mentioned earlier, the results will be presented again in line charts as visual support for the ease of reference during the discussion.

8.7.1 LP-TC interaction effects on syntactic complexity

The descriptive analysis with respect to the three measures of syntactic complexity showed that the +IR task generated more complex performance than the -IR task in terms of *mean length of AS-units* and *ratio of subordination* but it had no effect on *mean length of clauses*. B2 group produced the longest AS-units and the highest ratio of subordination while C1 group produced the longest clauses in the +IR task. However, A2 group produced the least syntactically complex performance in the two tasks. Figure 16 below recaps the performance of the four LP groups on the two tasks as captured by the three measures of syntactic complexity.

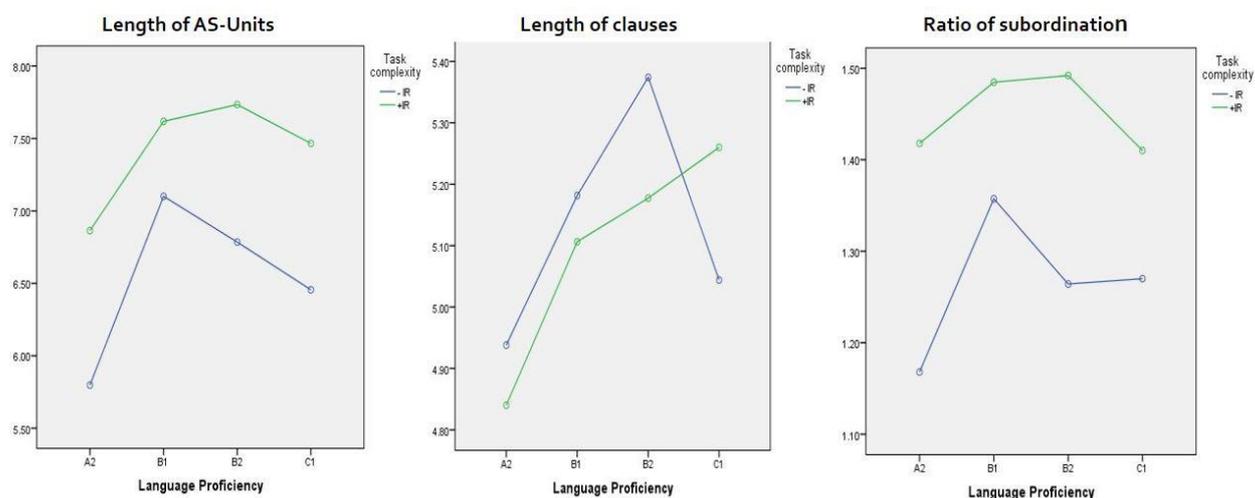


Figure 12. Effects of LP and TC on syntactic complexity

The line charts with respect to AS-units and subordination supported Norris and Ortega's (2009) claims that syntactic complexity keeps enhancing to a certain proficiency level, i.e. upper intermediate before it plateaus at higher proficiency levels. It was possible that more proficient learners started prioritising other aspects of language performance on complexifying the structures of their speech. It was also apparent that the nature of the task, i.e. a controlled oral narrative, restricted their performance, and therefore certain aspects of performance (e.g. syntactic complexity) may perhaps not go above a certain level.

As for the main effects of the two independent variables and their interaction effect on syntactic complexity, the two-way ANOVAs revealed only main effects for TC on length of As-units and ratio of subordination. However, the results did not show any main effects for LP on any of the three measures of syntactic complexity. Therefore, no interaction effect between LP and TC on syntactic complexity was established. These results confirmed that the performance varied in terms of syntactic complexity on the two tasks regardless of learners' level of LP. The linguistic requirements of intentionality and the choice of syntactic complexity measurements could be additional factors that diminished the effects of LP on the participants' output. The formulaic nature of the language which was found across all LP groups affected directly the length of As-units and clauses and therefore inflated the results regarding subordination. We assumed here that IR demands in this study as triggered by the differences between -IR and +IR task instructions and content did not interact differently with the participants' LP levels to affect syntactic complexity. As a result, no significant impact could be found or predicted for LP within the aforementioned factors.

8.7.2 LP-TC interaction effects on lexical complexity

As shown in Figure 17, the four LP groups produced more varied and sophisticated language in the -IR which contradicted the predictions that the more complex task (+IR) would generate more complex lexis. The line charts below recall that lexical diversity improved increasingly across the four LP levels, with the higher LP groups producing more diverse language and the lower LP groups producing less diverse lexis. It seemed that higher proficient speakers prioritised using more diverse lexis on using more complex structure. The results of lexical sophistication did not span nicely over the different LP levels. Still, the higher LP groups, i.e. B2 in the -IR task and C1 group in the +IR task used more infrequent lexis than the lower LP groups. It was possible that the small number of participants in C1 group, i.e. 5 learners influenced the results compared to 19 learners in B2 group. The results of lexical complexity offered confirmatory evidence that the less restricted requirements of the -IR task, i.e. *tell and describe the story* gave the participants more freedom to use more varied and infrequent vocabulary in the -IR task as captured by the measures of lexis. This was not available for the same participants in the +IR task which directed them to use specific linguistic performance to respond to the task-inherent and task-induced demands. The findings indicated that the higher LP groups, i.e. B2 and C1 responded more effectively than the lower LP groups to the linguistic demands in terms of choice of lexis due to their greater lexical repertoire.

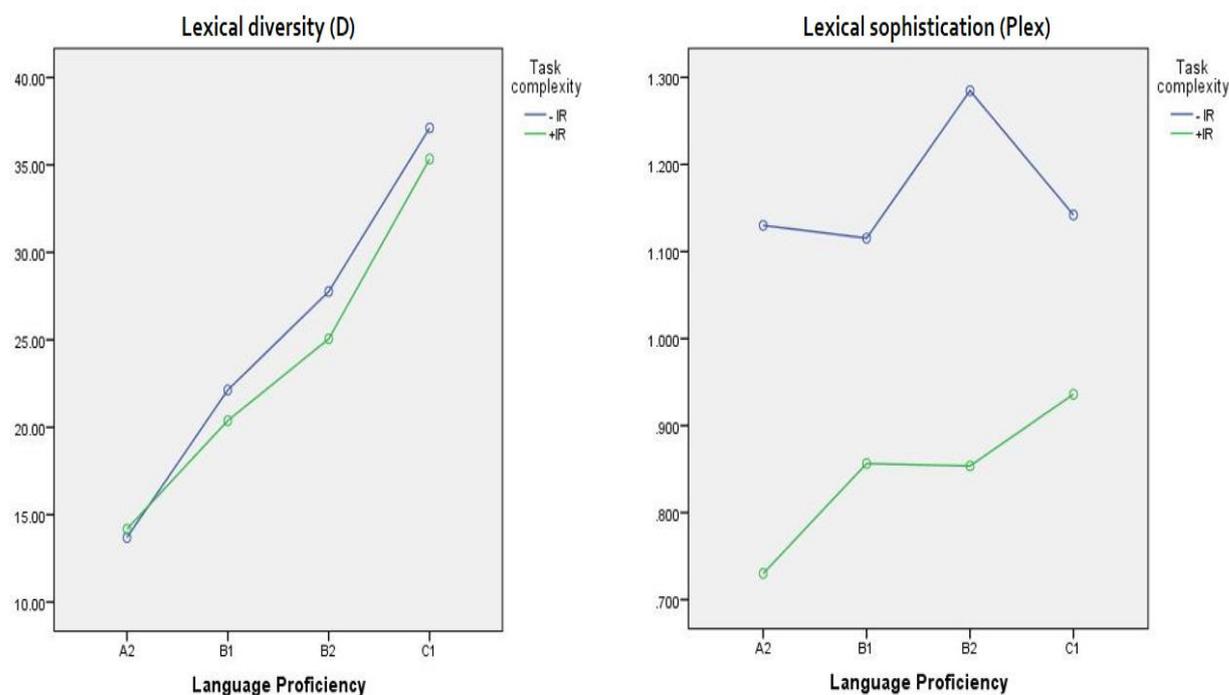


Figure 13. Effects of LP and TC on lexical complexity

The ANOVA results revealed a main and large effect for LP on lexical diversity. Another main effect was found for TC on lexical sophistication. However, no interaction effects between TC and LP were detected for both measures of lexical complexity. Based on the results obtained, it seemed fair to propose that the effect of varying degree of TC on lexical complexity was not moderated by varying learners' LP. It was therefore reasonable to conclude that increasing TC in this study influenced lexical complexity negatively regardless of the learners' level of LP.

8.7.3 LP-TC interaction effects on accuracy

Moving to accuracy, the descriptive analyses revealed that the performance of the four LP groups improved steadily in the two tasks. Regarding the effects of TC on accuracy, it was found that more accurate language was found in the +IR task for all LP groups. As depicted in Figure 18, the higher LP groups produced more accurate language than the lower LP groups as gained by measuring the percentage of error-free clauses and weighted clause ratio. As expected, it was found that as the level of LP increased, the accuracy of the participants' language performance increased. The more accurate language found in the +IR task across the different LP levels signified that the +IR demands were effective in directing the attention of the participants in all LP groups towards form at the expense of meaning. This indicated that regardless of LP level, the participants in this study prioritised attending form when the task increased in complexity as captured by the two measures of accuracy, i.e. *percentage of error-free clauses* and *weighted clause ratio*.

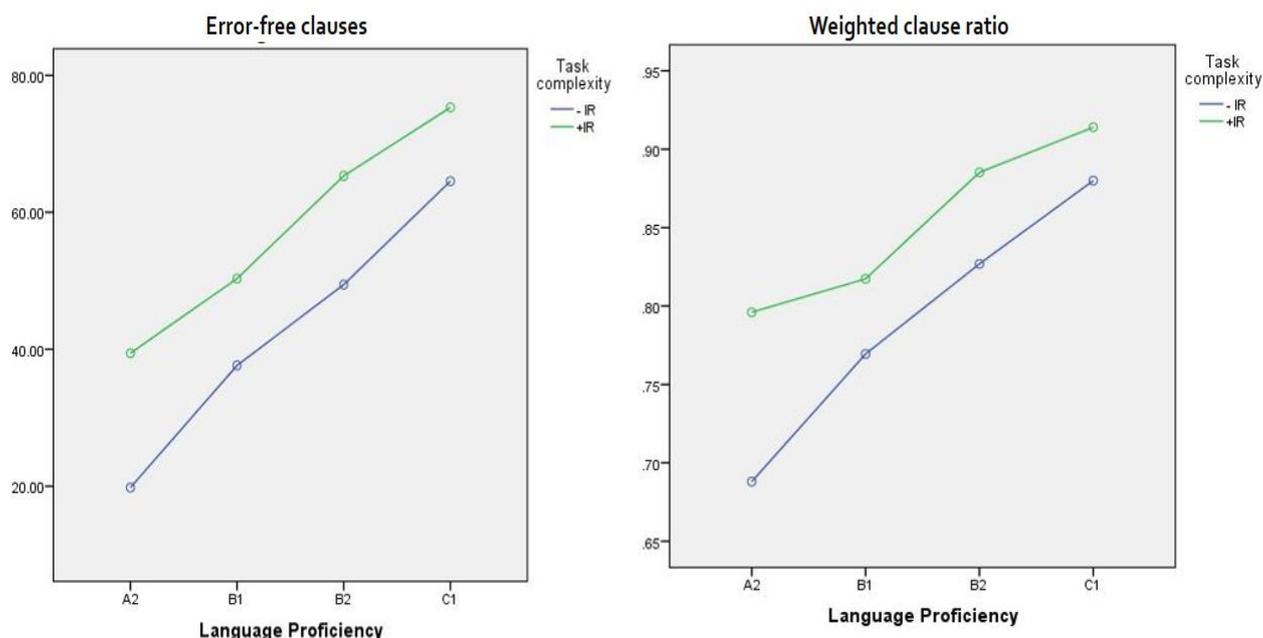


Figure 14. Effects of LP and TC on accuracy

The ANOVA outputs revealed main and large effects for LP and TC on the two measures of accuracy. In spite of the main effects of the two independent variables on accuracy, the interaction effects between LP and TC were not found to be significant with respect to *error-free clauses* or *weighted clause ratio*. Consequently, this would confirm that the positive effects of TC on accuracy were not moderated by differences in the participants' levels of LP and that TC affected all the LP groups equally. However, it was possible that the TC demands were not effective enough to direct the four LP groups to focus on accuracy differently. Another explanation for the nonsignificant interaction between LP and TC could be due to the fact that the variations in LP between each group were not large enough. Thus, achieving interaction effects would be viable if the participants were grouped into two groups, i.e. low versus high. Yet, maybe this is how different LPs act anyway, i.e. accuracy improves gradually but steadily.

8.7.4 LP-TC interaction effects on fluency

Turning to fluency, mixed findings were found with respect to the three facets of fluency, i.e. speed, breakdown, and repair. As shown in Figure 19 below, the results obtained from the two measures of speed fluency revealed that unpruned and pruned speech rates increased consistently across all LP levels in both tasks. Again, these nicely shaped charts suggesting that TC did not have an impact as moderated by LP. The higher LP groups produced higher speech rates and the lower LP groups produced lower speech rates in both tasks. Comparing between speed fluency in the -IR condition versus +IR condition, the higher unpruned and pruned speech rates were in favour of the +IR across all LP groups.

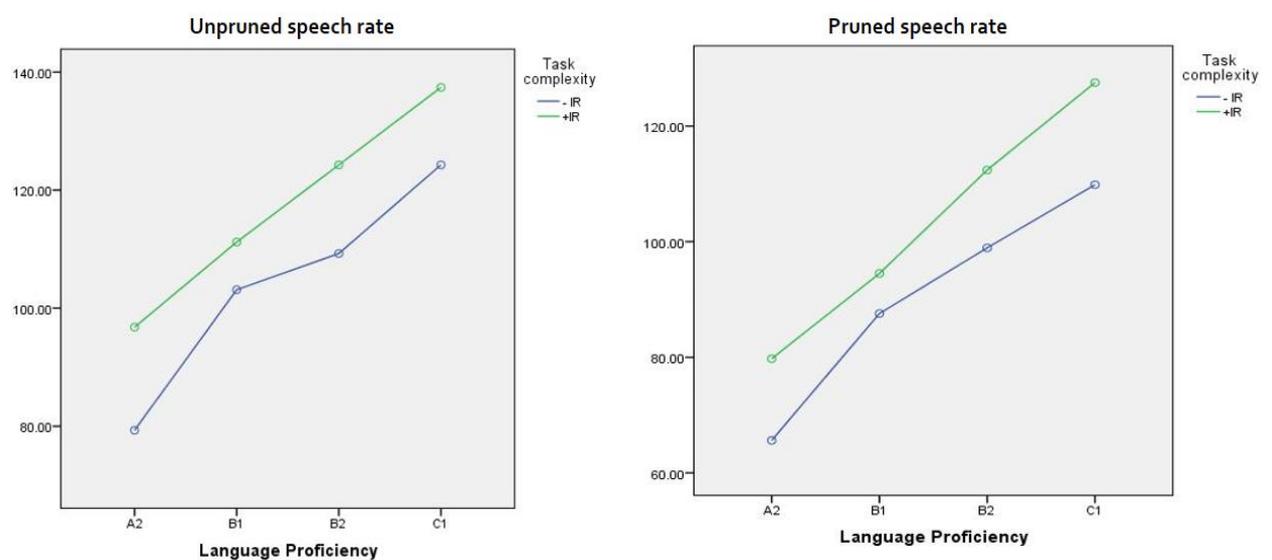


Figure 15. Effects of LP and TC on speed fluency

These results were against what was hypothesised that +IR demands would affect speed fluency negatively. It was apparent that the learners with higher LP benefited from the larger linguistic repertoire to speak faster. However, these advantages in terms of the size of the linguistic knowledge and the speed in accessing and retrieving lexis were not available for lower proficiency groups resulting in lower speech rates. Main effects were obtained for LP in both measures of speed but no effects were found for TC. Therefore, the interaction effects between LP and TC was not significant. The non-significant LP-TC combined effect on speed fluency suggested that the participants' speech rates enhanced when performing more complex tasks regardless of variation in their LP.

Moving to discuss the results of breakdown and repair fluency, only the measure of number of filled pauses achieved significant results, whereas all results of the measures of silent pauses and total number of repairs were not significant. The +IR condition encouraged more filled pauses than the -IR condition which was in line of the predictions. The results of filled pauses as depicted in Figure 20 below, showed that number of filled pauses declined consistently across all LP levels. The results indicated that as LP level increased, the number of filled pauses decreased with the highest LP group, i.e. C1 produced the least number of filled pauses. However, C1 group was the only group to produce the same number of filled pauses in the two tasks. This result could be attributed again to the small number of C1 learners. In this case, the results could be inflated if two participants were not doing what they were requested to do. The two-way ANOVA identified that the number of filled pauses was mainly affected by LP not TC as only a main effect was detected for LP. However, no interaction effect was confirmed for LP and TC on filled pausing. Therefore, it could be concluded that the variation in the number of filled pauses between the participants in the two tasks could not be attributed to a joined effect of variation in LP levels and TC demands.

As regards number of repairs, inconsistent patterns were found in both tasks. As shown below in Figure 20, B2 group made the lowest number of repairs in the two tasks. Interestingly, the lower LP groups, i.e. A2 and B1 produced the highest number of repairs in the +IR task. It was apparent that the IR requirements as triggered by +IR task instructions and content pushed the participants with lower LP to generate more repairs as a result of the increased cognitive TC. Nevertheless, none of the ANOVAs results in terms of the main and joint effects for LP and TC on repair fluency were significant. These findings suggested that the participants' behaviours regarding repairs (*e.g. repetitions, reformulations, replacements*) were not

associated with either TC manipulation or variation in the learners' LP as confirmed in previous research (e.g. Declerck and Kormos, 2012; Gilabert, 2007)

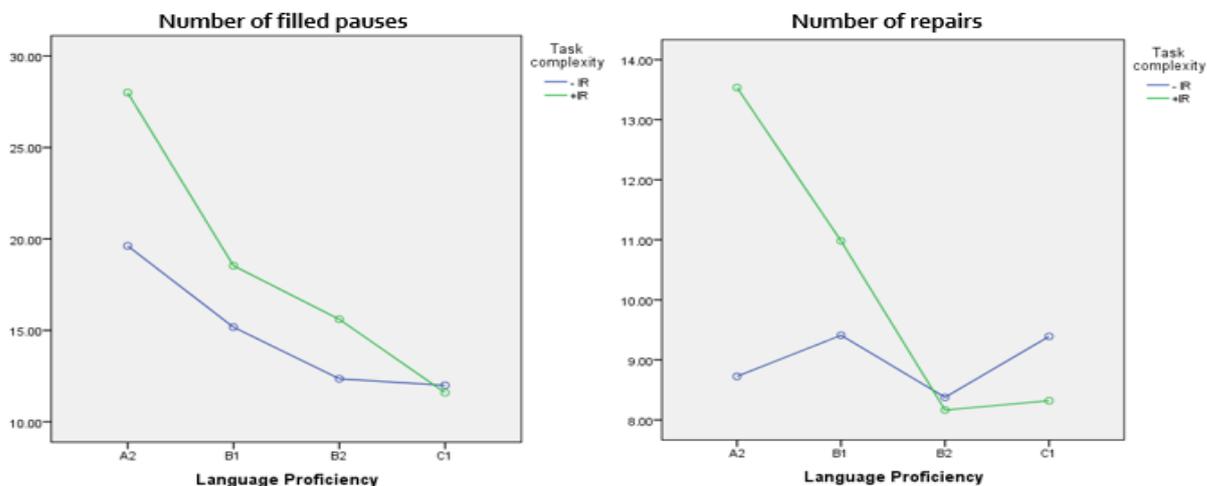


Figure 16. Effects of LP and TC on filled pauses and repair

8.8 WM and TC interaction effects on L2 performance

Research Question 4 was formed to examine whether variation in the learners' WM capacity moderated the effects of TC on the aspects of L2 speech performance, i.e. syntactic complexity, lexical complexity, accuracy, and fluency. Two-way ANOVAs were run with WM as a between-participants independent variable with three levels (low, medium, high) and TC as a within-participants independent variable with two levels (-IR, +IR). Measures of syntactic complexity, lexical complexity, accuracy and fluency were employed as dependent variables. It was predicted that higher WM groups would produce higher accuracy, lexical complexity, and speed fluency. However, no predictions were offered on whether the effects of TC would be significantly mediated by variation in WM.

The results confirmed only one main effect for WM on accuracy, whereas main effects were found for TC on certain measures of syntactic complexity, lexical complexity, accuracy and speed fluency. However, no interaction effects were detected for WM and TC on any aspect of L2 speech performance. These results pinpointed that the effects of increasing TC on L2 learners' oral performance functioned independently from any variation in the learners' WM capacity and that WM had no mediation effects on L2 speech performance. The following subsections discuss the findings in more detail with regards to the effects of WM and TC on each dimension of L2 oral performance, i.e. syntactic complexity, lexical complexity, accuracy, and fluency.

8.8.1 WM-TC interaction effects on syntactic complexity

The results of the three WM groups with respect to syntactic complexity revealed mixed patterns as shown in Figure 21. All WM groups produced enhanced syntactic complexity in the +IR task in terms of AS-units and subordination which was similar to the findings obtained by the LP groups. It was the medium WM group who produced the longest AS-units and clauses in both tasks and the highest ratio of subordination in the +IR task. These results could be understood since the medium WM group had more participants (N = 25) than the high WM group (N = 7). It was possible that the medium WM group responded more effectively to the IR linguistic requirements imposed by the instructions and content of the tasks which resulted in longer AS-units and higher ratio of subordination.

Surprisingly, the shortest clauses in the two tasks were produced by the high WM group. This result could indicate that the learners with high WM who were privileged to access and retrieve lexis faster than lower WM groups (Gilabert and Munoz, 2010) produced more formulaic sequences which resulted in shorter clauses. Consequently, higher ratio of subordination was also found in the language performance of the high WM group in the -IR task. This would posit that learners with higher WM would benefit from an easier and faster access to their mental lexicon to retrieve more formulaic chunks to respond to the IR demands. Thus, resulting in inflating their performance in terms of syntactic complexity which could mislead researchers in interpreting their findings.

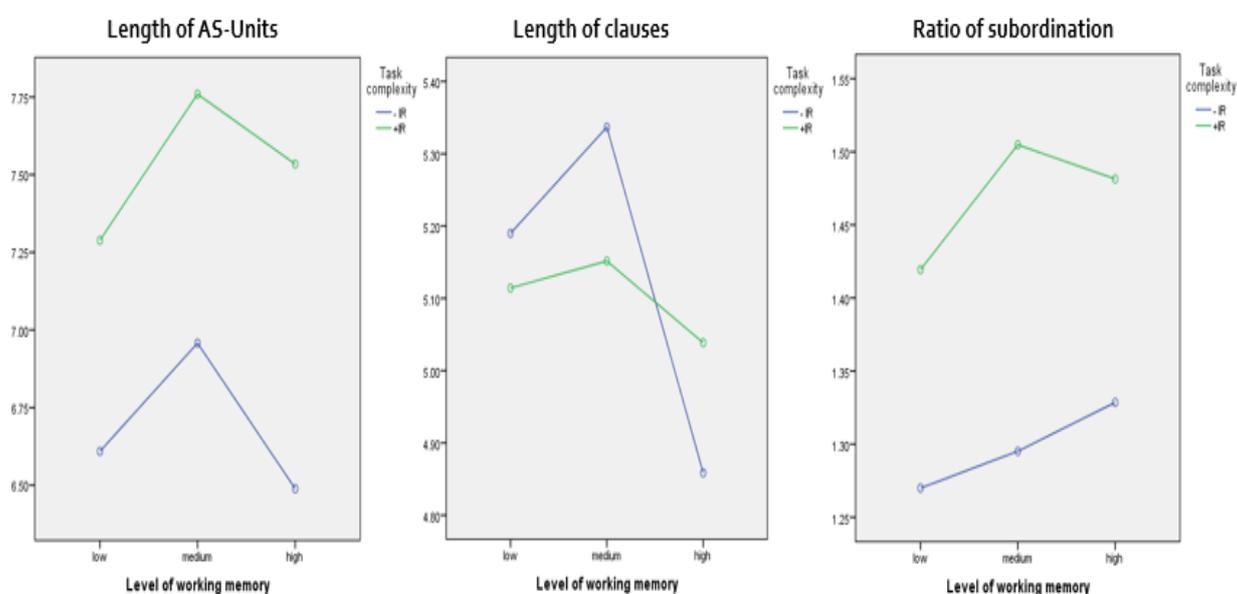


Figure 17. Effects of WM and TC on syntactic complexity

The interaction effect between WM and TC on syntactic complexity was examined by running two-way ANOVAs which revealed one statistical significant main effect for TC on mean length of AS-units. However, no main effects for WM were found on any measures of syntactic complexity. Therefore, no joint effect of WM and TC could be detected on syntactic complexity. These results suggested that the language performance in terms of syntactic complexity was not affected or moderated by differences in WM. These results contradicted those obtained by Kormos and Trebits (2011) who found that WM helped syntactic complexity but only in the less complex task. However, the results confirmed those found by Gilabert and Munoz (2010) who concluded that syntactic complexity did not benefit from WM on more complex tasks. Consequently, the predictions that WM interacted with TC to affect syntactic complexity were not confirmed.

8.8.2 WM-TC interaction effects on lexical complexity

The performance of the three WM groups as shown in Figure 22, revealed that the language produced in the -IR task was lexically more complex than the +IR task. Clearer patterns were noticed with respect to lexical diversity and sophistication in the -IR task with higher WM groups producing more varied and sophisticated lexis than the low WM group. It was possible that learners with higher WM benefited from their faster and better ability to access and retrieve lexis (Gilabert and Munoz, 2010) in a task that required no IR which was not available for learners with low WM in both tasks. However, mixed patterns were identified in the +IR task as the high WM group produced the least sophisticated lexis and the low WM group generated the most sophisticated language. One explanation to these results could be related to the way this study manipulated TC, i.e. IR inherent and induced demands which could diminish the impact of WM in the +IR task as it might have restricted the participants' choice of lexis. Another probable reason could be due to the small number of participants in the high WM group (N = 7) which might affect the comparability of the overall results.

The ANOVA results detected only one main effect for TC on lexical sophistication. The results revealed that WM had no main or joint effects on lexical diversity and sophistication. Respectively, it was not confirmed that the impact of TC on lexical complexity was mediated by variation in learners' WM. These findings did not lend support to Gilabert and Munoz (2010) who found that the performance with respect to lexical diversity was influenced by variation in WM due to the advantages high WM learners possessed regarding accessing and retrieving the required lexis more quickly than learners with low WM. The inconsistent findings could be

attributed to the differences between the two studies as Gilabert and Munoz’s study was correlational and was built on data that were gathered from only the complex task, i.e. film retelling task. It was plausible that the lexical requirements as imposed by the novel framework of this study did not allow the participants the opportunity to benefit for variation in their WM.

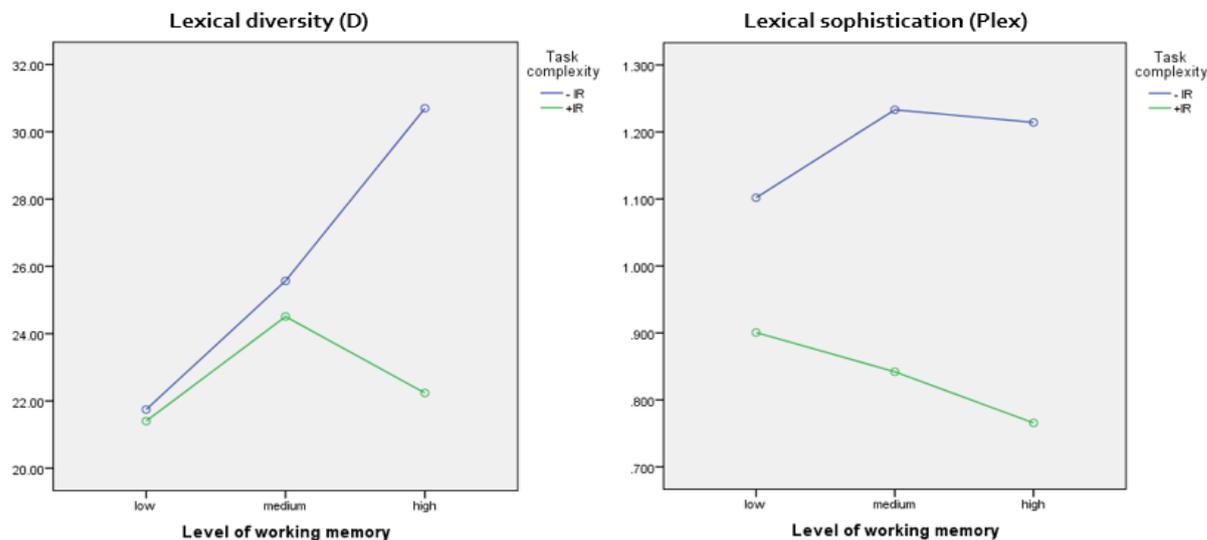


Figure 18. Effects of WM and TC on lexical complexity

8.8.3 WM-TC interaction effects on accuracy

Moving to accuracy, Figure 23 shows that the performance in terms of the two accuracy measures improved steadily across the three WM groups with higher WM groups (medium and high) producing higher error-free clauses and weighted clause ratio. All WM groups produced more accurate language in the +IR task. However, participants in the low WM (N = 16) produced the least accurate oral performance in the two tasks. It was possible that the medium correlation between WM and LP ($r = .471, p = .002$) contributed in these results. Still, these initial findings would suggest an effect of WM on the participants’ speech performance with respect to accuracy.

Running two-way ANOVAs showed significant main effects for WM and TC on the two measures of accuracy, i.e. *percentage of error-free clauses* and *weighted clause ratio*. Though, these effects might give an indication of a combined effect for WM and TC on accuracy, the ANOVA confirmed the WM-TC interaction effects on accuracy were not significant. This would suggest that the impact of TC on the accuracy of performance was not moderated by levels of WM. It could be apparent that the participants were relying more on their LP

knowledge than their WM while attending form. Consequently, accuracy enhanced across the tasks regardless of the learners' WM capacity.

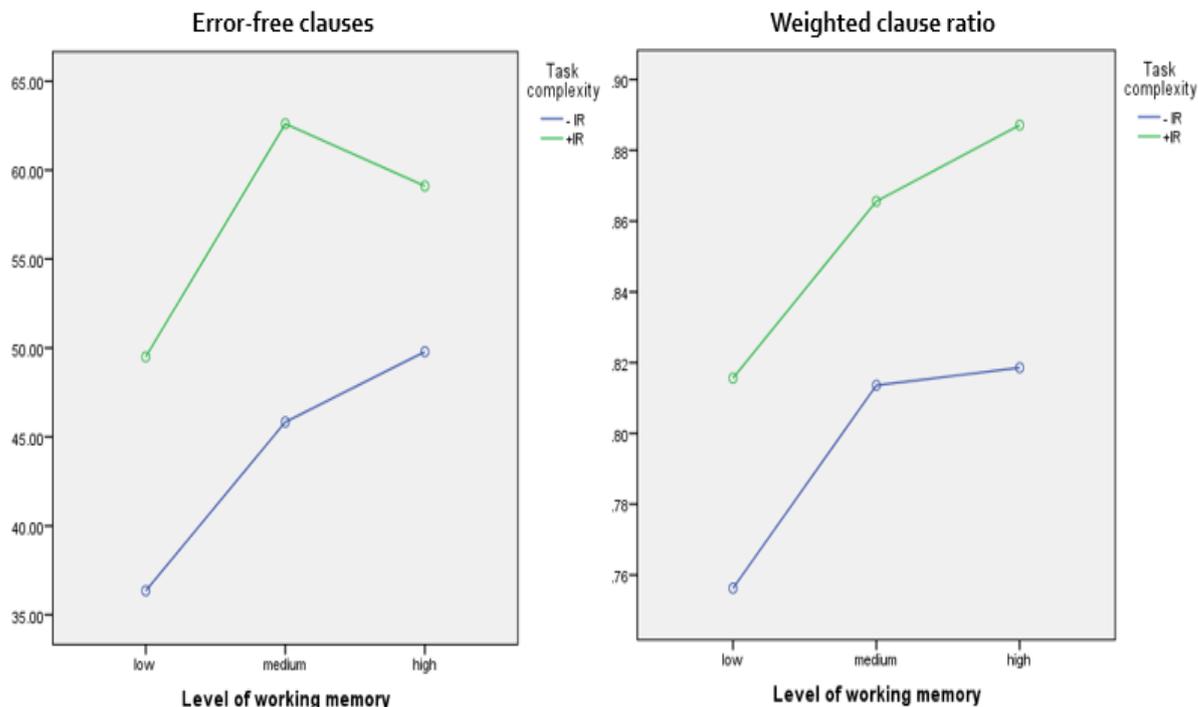


Figure 19. Effects of WM and TC on accuracy

8.8.4 WM-TC interaction effects on fluency

Concerning fluency, the measures of speed fluency, filled pauses and repairs will be discussed here. As shown in Figure 24 below, performance regarding speed fluency as captured by two measures of *pruned and unpruned speech rates* were in favour of the medium WM group which produced the highest speech rates in the two tasks. However, the low WM group produced the least speech rates in both tasks. It is worth mentioning that speed fluency was higher in the +IR task for the three WM groups. It was reasonable to find out that the low WM group produced low speech rate, but it was surprising to see that the medium WM not the high WM group produced higher speech rates in both tasks. The number of participants in the medium WM group (N = 25) compared to 7 participants in the high WM group might have played a role in these results. However, these results were still supportive to the claims that high WM could be beneficial in promoting speed fluency (Gilabert and Munoz, 2010).

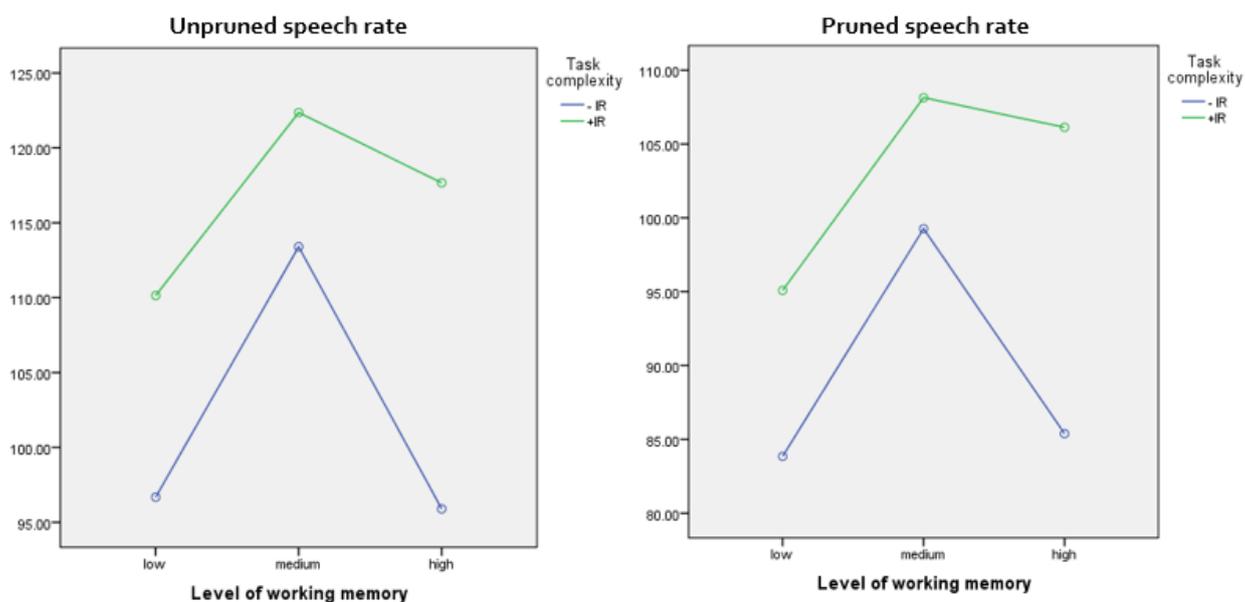


Figure 20. Effects of WM and TC on speed fluency

However, the ANOVA results failed to confirm any main effects for WM on any measures of speed fluency. The only significant main effect was in favour of TC on *unpruned speech rate*. Hence, no interaction effects between WM and TC on speed fluency could be established. These findings rejected any predictions with respect to the combined impact of WM and TC on speed fluency and suggested that TC as operationalised in this study had a positive effect on speed fluency regardless of the learners' WM capacity. These findings did not support other studies which found that WM had positive impact on speed fluency as a result of its positive effects on lexical complexity (e.g., Gilabert and Munoz, 2010).

Turning to discuss the effects of WM and TC on filled pauses and repairs, Figure 25 below illustrated that the high WM group produced the lowest number of filled pauses and repairs in both tasks (-IR and +IR). It was the medium WM group which made more filled pauses and repairs than the other groups. Again, these results could be attributed to the small number of high WM participants ($N = 7$). It could be also possible that the learners with high WM felt that they did not need to use many filled pauses or do repairs more frequently benefitting from their extended WM capacity resulting in positive gains on their fluency. The ANOVAs results failed to detect any significant main effects for neither WM nor TC on filled pauses or repairs. As a result, no WM-TC joint effects were found on the learners' total number of filled pauses or repairs. Following these findings, it could be concluded that the impact of TC on filled pauses and repairs was independent from any moderation of the learners' WM capacity.

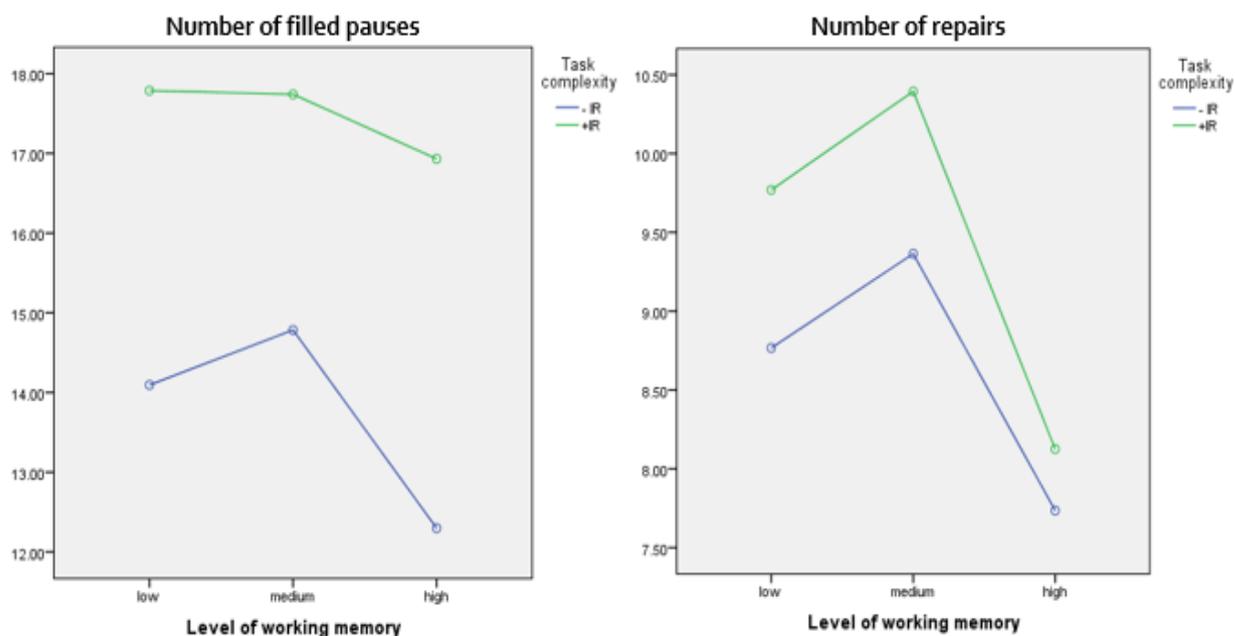


Figure 21. Effects of WM and TC on filled pauses and repairs

8.9 LP and WM predictability of L2 oral performance

Research Question 5 was designed to examine the power of LP and WM in predicting the learners' L2 oral performance as captured through composite measures of syntactic complexity, lexical complexity, accuracy, speed fluency and pausing fluency. It was expected that LP and WM would contribute to explain variations in the participants' speech performance regarding lexical complexity, accuracy and speed fluency but not syntactic complexity or pausing fluency. The moderate correlation between LP and WM ($r = .471$) suggested that each independent variable would predict different unique features in each performance aspect and would form together successful regression models. Running standard multiple regression analyses revealed significant regression models with LP and WM as predictors of learners' speech performance regarding lexical complexity, accuracy, speed fluency, and pausing fluency but not syntactic complexity.

As predicted for syntactic complexity, the nonsignificant regression model suggested that LP and WM were not reliable predictors of L2 performance regarding syntactic complexity. These findings were supportive to the claims that L2 learners tended to ignore complexifying their grammatical structures when they attained higher levels of proficiency (Norris and Ortega, 2009). Moreover, WM was assumed to facilitate L2 learners' speech performance in terms of using complex lexis but not complex structure (Gilabert and Munoz, 2010). Hence, it was

concluded that the variations between L2 learners in their syntactic complexity were not explained through variations in their LP and WM.

As regards lexical complexity, the significant regression model confirmed that LP and WM together explained 38.2% of the variation in the learners' lexical complexity in terms of diversity and sophistication. The strong correlation between LP and lexical complexity ($r = .62$) and the significant individual contribution to the model ($p = .000$) proposed LP as a stronger contributor to explain variations regarding lexical diversity and sophistication compared to WM which failed to achieve significant contribution despite its small correlation with lexis ($r = .26$). It was therefore feasible to claim that learners with higher LP were likely to produce more varied and sophisticated lexical items even if their WM capacity was not high. These findings were in line with Gilabert and Munoz (2010) who advocated LP as a stronger predictor of lexical complexity than WM. It was apparent that the participants were relying more on their L2 knowledge and proficiency rather than their WM in accessing and retrieving the required lexis. The linguistic requirements of the two tasks might not be difficult enough to tax the participants' WM, and hence its capacity was not a crucial factor in maintaining more varied and sophisticated lexis.

Turning to accuracy, the significant regression model suggested that this model is more successful than the other models in predicting the participants' speech performance by explaining 52% of the variance in the *percentage of error-free clauses* and *weighted clause ratio*. The high correlation between LP and accuracy ($r = .72$) and the medium correlation between WM and accuracy ($r = .35$) suggested the two independent variables as unique contributors in predicting the level of accuracy in L2 learners' performance based on tasks with increased TC. However, a significant individual contribution was only spotted for LP ($p = .000$) but not for WM. Hence, the results confirmed LP as a stronger contributor in predicting accuracy than WM. It seemed that the effects of manipulating IR through task-induced and task-inherent demands were sufficient to direct the learners' attentional resources towards accuracy more than any other aspect, and thus activate their form schemata resulting in more predictable language performance in terms of accuracy.

Moving to fluency, a significant regression model was also found regarding speed fluency suggesting that LP and WM together had the power to predict only 23% of variation in terms of unpruned and pruned speech rates. A medium correlation was found between LP and speed fluency ($r = .48$), but surprisingly, negligible and nonsignificant correlation was detected

between WM and speed fluency ($r = .17$). Moreover, the individual contribution of each variable to the model was significant for LP ($p = .001$) but not for WM, and thus promoting only LP as a predictor of speed fluency. Finally, a regression model tested LP and WM predictability of speech performance in terms of pausing fluency measured by *number of filled* and *silent pauses*. The model also revealed that LP and WM together explained only 17% of the variance. Again, only LP correlated with pausing fluency ($r = .41$) which also contributed significantly to the model but it was not the same case for WM. These findings would support what was observed for speed fluency that LP and WM together were not reliable variables to explain variation between L2 learners in their fluency.

LP appeared as a stronger variable than WM in predicting L2 speech fluency with respect to speed and pausing. These findings would indicate that variations in speed fluency could not be explained through the participants' variation in WM capacity. This conclusion opposed Gilabert and Munoz (2010) who found a correlation between WM and unpruned speech. These inconsistent results could be attributed to the way Gilabert and Munoz measured WM, i.e. reading span test and the way they operationalised TC, i.e. film retelling. Furthermore, the findings partially contradicted those obtained in Mota (2003), who observed small to medium correlations between WM, measured by speaking span tasks, and speed fluency but not pausing fluency. However, the findings supported other studies which also found that speech fluency could not be predicted based on learners' variation in WM (e.g., Kormos and Trebits, 2011).

8.10 Conclusion

This chapter discussed the results of Study Two which were obtained to answer the five research questions. The findings were explained and discussed in more detail with regards to the hypotheses, TC models, speech production models, and previous research on TC and individual differences in language proficiency and working memory. The next chapter will present the final remarks from Study One and Study Two. The implications and contributions of the two studies for L2 pedagogy and research will be then highlighted. The limitations of both studies will be mentioned and suggestions for future research will be finally made.

Chapter 9: CONCLUSION

9.1 Introduction

This closing chapter draws conclusions from the main findings of Study One and Two. Then, it highlights the significance of the achieved results and their contributions to advancing research in Task Complexity (TC) and Task-Based Language Teaching (TBLT). The chapter then discusses the implications of the findings for research and pedagogy in the fields of TBLT and SLA. The limitations of the study are then acknowledged. The chapter concludes with proposing some potential areas for future research in TC, TBLT and SLA.

9.2 Conclusions from the findings

This thesis reported on the findings obtained from two interrelated studies that were motivated chiefly by a need for a more systematic and in-depth investigation of intentional reasoning (IR) as a Task Complexity (TC) factor. Study One was designed to examine the effects of increasing TC along IR requirements on, 1) L2 learners' speech performance; 2) and perceptions of Task Difficulty (TD). Following the findings of the first study, a second study was further designed to investigate, 1) whether the effects of TC as operationalised by IR demands on L2 speech performance interact with learners' variation in language proficiency (LP) and working memory (WM); and 2) whether learners' performance on tasks that require varying degrees of IR can be predicted through the learners' LP and WM. The participants in the two studies performed two video-based oral narrative tasks with varying degrees TC, i.e. -IR and +IR and their speech performance was recoded, transcribed and analysed against several measures of syntactic complexity, accuracy, lexical complexity and fluency (CALF). They then completed a retrospective questionnaire on how and why they perceived each task in terms of difficulty. IR was operationalised at two levels, i.e. task instructions and task content to ensure a more thorough investigation of this construct.

9.2.1 Conclusions from Study One

With a means of running paired-sample t-tests, the findings of Study One indicated that the more complex task (+IR) produced more accurate and syntactically complex language than the less complex task (-IR). The latter surprisingly elicited more lexically complex oral performance, whereas no significant effect was detected on fluency. The findings lent mixed support to Robinson's (2001) Cognition Hypothesis (CH), which posits that performing more cognitively complex tasks is assumed to be associated with more accurate and complex but less

fluent language performance. The predictions of the CH are confirmed with regard to accuracy and syntactic complexity but were refuted with regard to lexical complexity, whereas the predicted less fluent performance due to the increased IR demands was not confirmed.

The findings with respect to TD revealed that the participants perceived the more complex task (+IR) as more difficult than the less complex task (-IR) as the CH predicted. The qualitative analysis of the participants' justification of their judgement of TD showed that the participants mainly attributed TD to, 1) task-induced cognitive demands as imposed by task instructions; and 2) task-inherent cognitive demands as imposed by task content. These results further supported the validity of the choice of the video tasks and confirmed the successful manipulation of TC through offering a two-level framework to operationalise IR. Despite this successful operationalisation of IR demands in Study One, the inconsistent findings in terms of the effects of TC on L2 speech performance opened new areas of enquiry, which suggested that task factors were not the sole contributor to task performance. Study Two was therefore pursued to address this enquiry by exploring the role of learner factors, i.e. LP and WM in performing tasks with increased TC.

9.2.2 Conclusions from Study Two

The findings of Study Two confirmed those obtained in Study One with respect to the effects of IR on L2 learners' speech performance. As attained by running paired-sample t-tests, performance on the +IR task was characterised with increased syntactic complexity, accuracy, speed fluency, and filled pauses. However, the +IR demands resulted in decreased lexical complexity, while no effect was found on silent pausing or repair fluency. These results again lent support to the CH in terms of the positive effects on syntactic complexity and accuracy, but pointed in the opposite direction regarding lexical complexity. In terms of the predictions of the CH about the negative effects of TC on fluency, mixed findings were found. The increased number of filled pauses in the +IR task backed the CH, whereas the higher speech rates in the same task contradicted the predictions. However, the results of silent pausing and repair fluency were mixed and not significant, which matched what was achieved in Study One.

The findings of TD retrospective questionnaire in Study Two were fully consistent with those found in Study One. The participants again rated the +IR task as more difficult than the -IR task which lent further support to the predictions of the CH. The thematic analysis revealed the similar sources that contributed to perceptions of TD. Receiving 84% of the responses, both

task-induced and *task-inherent* cognitive demands were confirmed as the key factors that influence the participants' perceptions of TD, whereas, *linguistic demands* and *time pressure* received 16% of the comments. It was apparent from the qualitative analysis that the IR requirements as expressed in the +IR task instructions (e.g., predict behaviours, read thoughts, explain intentions, and justify actions and reactions), in addition to the content of the +IR video (e.g., unfamiliar, unpredictable, unclear) had a direct impact on the participant's rating of TD. These results further confirmed the validity of the framework this study put forward to operationalise IR more systematically at the levels of task instructions and content.

Respecting learners' individual differences, the findings of Study Two with respect to the interaction effects of LP-TC and WM-TC on speech performance failed to achieve significant levels as obtained through running two-way ANOVAs. Only individual effects were detected for each independent variable (LP, WM, TC) on several CALF measures, but no interaction effects between LP and TC or WM and TC could be found in Study Two. These findings suggested that the effects of TC as operationalised in this study through IR demands on L2 speech performance were not mediated by variation in the participants' LP or WM.

The findings that were obtained by a means of multiple regression analyses to see whether LP and WM had the power to predict speech performance of tasks with increased TC revealed mixed findings. Only the regression model regarding syntactic complexity was not found to be significant, which suggested that neither LP nor WM could explain the variations in the participants' syntactic complexity. However, the LP-WM regression models were found to be significant with respect to lexical complexity, accuracy, speed fluency, and pausing fluency. Furthermore, LP and WM together were found to explain 17% to 52% of the variations in the participants' speech performance with the accuracy model as the most successful one. However, only LP was found to have significant individual contributions to these models, whereas WM failed to contribute significantly to all regression models. However, the significant but small correlations between WM and the composite measures of accuracy and lexical complexity offered indications that WM could predict performance pertaining to these two performance aspects but to a limited extent.

9.3 Contributions of this study

This study is designed to offer original contributions to research within TBLT and SLA at conceptual and methodological levels, and to address the issues that have emerged in previous

research. One of the strengths of this study is drawing on the cognitive psychology to redefine the construct of intentional reasoning (IR) to address the limitations of previous research in this area (e.g., Ishikawa, 2008; Robinson, 2007). As discussed in previous chapters, up to now IR had been inadequately defined and operationalised. A systematic approach to defining and operationalising IR, one of the key contributions of the current study, can offer guidance for future research to put forward more systematic and robust operationalisations of this construct. Therefore, this study proposes a more comprehensive definition (see Section 2.8) that embraces the different components of IR, i.e. *beliefs, desires, views, knowledge, intentions, actions, reactions, and consequences* (Bratman, 1987), and the steps of reasoning about other people's intentions, i.e. *observing actions, understanding beliefs and desires, drawing conclusions and predicting actions and their consequences* (Astington & Baird, 2005).

Responding to the inconsistent and unsystematic investigations of IR, this study proposes a novel framework to operationalise IR at the levels of task instructions and task content. The task instructions explicitly encourage the participants in the +IR task to read thoughts, identify desires, reason about intentions, justify behaviours, predict actions, and explain their consequences while telling and describing what is happening in the video. The instructions written in English and Arabic have been used to ensure that all participants understand what reasoning about other people's intentions entails or requires. Concerning task content, this study applies a careful and systematic procedure to select the video clips that are assigned to each task. The selection process adheres to De Jong and Vercellotti's (2016) framework of employing narratives in task-based studies that are comparable except in the construct(s) under investigation to ensure successful and systematic task manipulation.

The in-depth investigation of TD is another novel contribution this study attempts to offer. As it is expected that the participants will rate the +IR task as more difficult, this study further investigates carefully how and in what ways they will find it challenging. Thus, this thorough exploration of TD can inform the debate over the constructs of complexity and difficulty and how far they overlap/intersect or not. Combining the quantitative analysis of perceptions of TD with a careful qualitative investigation is meant to elucidate the links between IR, TD and TC, and hence contribute to establish an index of TD. As suggested by Révész et al. (2016), the thematic analyses of TD in Study One and Two have also served as a tool to validate the novel manipulation of IR at two levels, i.e. instructions (posing induced IR demands) and content (posing inherent IR demands). These results suggest that manipulating TC along task content

is not the only key contributor in shaping TD perceptions, but likewise manipulating TC along task instructions can influence both performance and perception. Furthermore, higher percentage of comments in Study Two have mentioned IR demands as triggered by task instructions (43%) as the main source of TD perceptions, compared to 41% of the comments that have declared IR demands as imposed by task content as the main contributor to TD.

Another methodological contribution of this study is addressing the issue of validating, checking and extending the CALF measures. Study One and Two have employed a wide range of measures that tap into the various facets of each aspect of speech performance. As the first study has employed 12 measures, the second study has further recruited twenty measures. The two studies have used the same three measures of syntactic complexity to tap into length and subordination (Norris & Ortega, 2009). Concerning measures of lexical complexity, Study Two has included a measure of sophistication to respond to the limitation of Study One which has employed only one measure of lexical diversity. Regarding measuring accuracy, the first study has adopted two measures that consider error count and error-free language. However, in response to the recent calls to consider measuring accuracy based on error gravity, Study Two has included *weighted clause ratio* (Foster and Wigglesworth, 2016), in addition to *percentage of error-free clauses*. Respecting fluency measures, the study suggests filled pauses as a more sensitive measure than measures of silent pausing in tapping in breakdown fluency. The results of the measures of speech rate need to be interpreted with more caution as they can be affected by task requirements for lexis. In this study, the more complex task encourages the use of more repetitive lexis of formulaic nature which contributed to increasing the participants' speech rates regardless of their level of LP or WM.

The findings of the two studies have revealed interesting interactions between speech performance and the choice of CALF measurements. The way this study operationalises IR at the levels of task instructions and content seems to have a direct impact on what CALF measures have captured. It is apparent that the IR induced and inherent requirements in the +IR task have stimulated and pushed the participants to produce language that is characterised with more formulaic sequences, mental state verbs, logical connectors, modal verbs, and adverbs of uncertainty. These findings have been further confirmed through running a post hoc frequency analysis of lexis. Consequently, the speech performance in the +IR task as captured by CALF measures incorporates certain patterns that are not frequently found in the -IR task (e.g., shorter clauses, longer AS-units, inflated subordination, and repetitive lexis).

These patterns explain why mixed and inflated results have been found with respect to syntactic complexity in this study and maybe in previous TC research (Malicka and Sasayama, 2017). These findings support the need to consider task essentialness for subordination (Inoue, 2016), and recommend interpreting the results of syntactic complexity with more caution. The features of language performance in the +IR task (e.g., formulaic chunks and repetitiveness of certain lexis) have also a direct effect on increasing accuracy and speed fluency, while simultaneously decreasing lexical complexity. These results indicate that IR demands as operationalised in this study have led to complexify structure at the expense of lexis. The conservative use of more diverse and sophisticated lexis and the task-content dependent lexis are the main characteristics of speech performance in terms of lexical complexity in the more complex task. The participants in the +IR task have used what is assumed to be more lexically diverse lexis (e.g., psychological state verbs, logical connectors, adverbs of uncertainty), but have produced a smaller *D* due to the lexis repetitiveness behaviour as a result of IR demands. Relating this performance behaviour to the way IR is operationalised, suggests a methodological confound, rather than a linguistic effect, which can be regarded as a novel contribution of this study with respect to the careful consideration of a possible interaction between manipulating the independent variables under inquiry and choice of CALF measures.

The last key contribution of this study worth highlighting here is the way language proficiency (LP) is thoroughly measured and operationalised. This study attempts to address the criticism different TBLT studies have received for using only one test to assess LP, and to ensure more in-depth operationalisation of this construct. Therefore, this study has employed a test that taps into L2 learners' explicit knowledge, i.e., Oxford Placement Test (OPT) and a test that measures learners' implicit knowledge i.e., Elicited Imitation Task (EIT) (R. Ellis, 2009b; Erlam, 2006). The decision to combine OPT with EIT to operationalise the learners' overall LP, is justified by an existing evidence that L2 explicit and implicit knowledge are distinctive constructs, and therefore they need to be tested by separate measures that are designed to tackle each construct to achieve thorough operationalisations of LP (R. Ellis, 2009b). Furthermore, Elder and R. Ellis (2009) claim that most of L2 written standardised tests are assumed to tap into explicit knowledge rather than implicit knowledge. The study here assumes that both explicit and implicit knowledge may be used even in online spontaneous speech performance, so ensuring both elements are pre-tested provides a particularly thorough framework for establishing LP as an independent variable for future TBLT-based research.

9.4 Implications of this study

This study sought to advance our knowledge of how manipulating TC along IR influences L2 speech performance in interaction with learners' variations in LP and WM. Therefore, this study offers implications for theory, methodology and pedagogy. The theoretical implications can be linked to the models of TC, i.e. the Cognition Hypothesis (CH) and the Limited Attentional Capacity (LAC). Given that testing the predictions of the CH is one of the aims of this study, the findings provide mixed support to this model which predicts positive effects of TC on complexity and accuracy at the expense of fluency. Only the +IR positive effects on syntactic complexity, accuracy and filled pausing are in harmony with the CH. The results of lexical complexity which are in contrast with the CH, point toward a possible trade-off between syntactic complexity and lexical complexity. Therefore, not only the predictions of the CH (complexity-accuracy jointly increase) are not confirmed, but also the LAC predictions (complexity-accuracy trade-off) are not supported. These mixed findings that support the recent meta-analyses for TC research (Jackson and Suethanpornkul, 2013; Malicka and Sasayama 2017) suggest a need to review how TC variables are being defined and investigated to harmonise the research efforts that explore TC factors.

The predictions of the CH regarding learner factors have also received variable support in this study. The findings with respect to perceptions of TD are in harmony with the predictions of the CH. However, the qualitative analyses have shown that task requirements (induced and inherent) are key contributors to the participants' judgement of TD. These results may imply that the way TC is manipulated may load on L2 learners' affective perceptions of TD regardless of their individual cognitive differences in WM. In terms of the predictions of the CH that learners' individual differences in their cognitive abilities (e.g. LP and WM) play key roles in speech performance on tasks with increased TC, no clear evidence can be found in this study to support these claims. Study Two has failed to detect any combined interaction effects between learner factors and TC. Only LP is confirmed as a reliable predictor of all aspects of speech performance except syntactic complexity, whereas WM effects on speech performance as a mediator or predictor are not confirmed.

This study offers novel implications in terms of how to systematically operationalise the constructs under investigation. This study highlights the importance of re-defining the main variables carefully before any operationalisation is proposed. Therefore, researchers need to ensure that their operationalisations of any variables under investigation in their studies are

systematic, validated, supported theoretically and methodologically, piloted carefully, and are taking into consideration the shortcomings of previous research. This applies to the way this study has re-operationalised IR through a two-level framework, which has been further justified and validated.

Another implication that can inform the methodology of TBLT and SLA research is adopting mixed-methods approaches to ensure in-depth and thorough exploration of the variables that are researched. Collecting retrospective quantitative and qualitative data from the participants about their performance and participation can provide valuable information to depict a better holistic picture about the area of investigation and to improve and validate the researchers' interventions. The last methodological implication is the choice of the measures to assess the variables of a study. The study has followed systematic procedures in selecting and implementing the measures of speech performance, LP and WM to ensure the validity and reliability of the obtained scores. All the measures employed in this study have been piloted carefully and their choice has been justified.

The study offers some important implication for L2 pedagogy, as it demonstrates that manipulating TC in L2 classrooms has a positive effect on facilitating focus on form, and thus promoting L2 performance. In addition to enhancing accuracy and grammatical complexity, manipulating pedagogic tasks as proposed in this study can encourage learners to utilise certain lexis (e.g., mental state verbs, connectors, adverbs of certainty). Consequently, these tasks provide learners with opportunities to use and acquire certain lexical units, and further help teachers predict the linguistic patterns to be produced by their learners. The observations from this study indicate that monologic oral tasks form a good platform for learners to build their self-confidence in speaking by challenging them to perform under time pressure without prior planning. Finally, the successful operationalisation of IR through task instructions and its significant impact on language performance, suggests that task instructions are key components of task design. Hence, it is important for syllabus designers and teachers to consider more careful designation of task instructions, taking into consideration the aim and the desired outcome of the task. The findings of this study can further inform decisions in terms of sequencing tasks with increased TC within syllabus design.

9.5 Limitations and suggestions for future research

A number of limitations can be acknowledged in this study and should be addressed in future research with regards to this study. This study has been implemented in a quasi-experimental setting outside the classroom context. Therefore, the findings could have limited pedagogical implications for L2 classrooms. It is necessary for future research to see if these results will be replicated when the study is conducted in classroom settings. Another limitation stems from the fact that the study investigates the effect of manipulating TC on monologic oral performance, and thereby the findings may not be generalisable to other task types, characteristics or modes. Future research should look into the effects of IR on for example dialogic tasks as the interaction between the interlocutors may change the effects of IR on performance. One of the limitations might be the use of only general measures to operationalise the aspects of speech performance, and therefore fails to capture a refined picture of IR effects, which can be captured by a means of specific measure. Future research should attempt to employ specific measures that can be more sensitive in capturing variation in language performance between tasks with increased demands or different requirements.

For reasons of practicality, it was not feasible to adopt a battery of WM tests to ensure more robust assessment of the participants' WM capacity. It is recommended that future research use computerised versions of WM tests rather than using one-to-one tests which are more time-consuming. This will help future studies save time and efforts, and overcome the practical issues that emerge during data gathering. Another practical issue has emerged regarding the low number of participants available in each of the lowest and highest LP groups, i.e. A2 and C1 (N=5). Future research is advised to recruit more participants to increase the possibility to allocate more participants to each group or employ binary levels of LP, i.e. low and high, instead of using four levels as in this study. By having more participants in each group or ensuring variations between learners' LP when employ binary level, the reliability of the findings will be boosted.

The findings of this study should be interpreted and generalised carefully. The results have been largely affected by, 1) the nature of the employed tasks (monologic video-based narratives); 2) the IR requirements (induced and inherent); and 3) the choice of CALF measures. This controlled task conditions, characteristics, and analysis may have led to control the participants' language performance, and thus requires more caution in explaining and generalising the findings. Future studies are advised to wisely consider the interaction between the

abovementioned elements to guarantee sufficient research planning and implementation. Further research is also encouraged to explore the interaction effect between learners' affective factors (e.g., anxiety and motivation) and IR demands on L2 speech performance, which is beyond the scope of this study.

Another area for further research is exploring the relationship between learners' perceptions of TD and their actual language performance. While it is evident that manipulating TC can have a direct impact on L2 learners' perceptions of TD, it is not clear whether their perceptions of TD can also have consequences on how the learners perform the tasks with increased TC. Given the fact that this is beyond the scope of this study, future studies are encouraged to investigate TD as an independent variable rather than a dependent variable. It is expected that responding to this research question can chiefly contribute to TBLT studies both theoretically and methodologically, and have valuable pedagogical implications for teaching and material design.

Research examining TC and its interaction with learners' cognitive ability factors is advised to ensure a reliable understanding of the constructs under investigation from broader perspectives. Therefore, researchers are advised to redefine the investigated variables and rebuild their operationalisation frameworks by visiting the neighbouring disciplines (e.g., cognitive psychology) to borrow definitions and theoretical framework to ensure more robust examination of these constructs. Finally, this study has operationalised TC through only a binary model, i.e. *less complex* versus *complex*. Future studies are advised to consider investigating cognitive complexity through applying a continuum of TC. However, researchers need to be aware that the designed tasks are comparable but different only in terms of TC demands. It is also recommended that future research explores the effect of task sequence, i.e. *complex-simple* versus *simple-complex* on L2 speech performance and the perceptions of TD.

This study recruited younger participants at the age of 16 to respond to the lack of TC studies that consider different age groups. The key role that age might play on tasks with increased TC is an area that has received little attention. Further studies examining younger learners who have not reached cognitive maturity are needed. Such studies can help understand whether variables such as TC, LP and WM interact with cognitive development and maturity, and may offer valuable pedagogical implications for TBLT and SLA. The lack of longitudinal studies that examine the effects of TC on language development is the final issue which needs highlighting. Further longitudinal research is required to investigate how TC can contribute to L2 development over a long period of time.

9.6 Final remarks

The complex picture that has been depicted in this study, will pose more challenges on the cognitive theories that form the backbone of TC research. The Cognition Hypothesis and the Limited Attentional Capacity are still required to gain more support for their claims and proposals from the current empirical research in SLA, LT and the neighbouring disciplines. Researchers that are interested in investigating TC are challenged to offer more collaborative efforts that help reach consensus regarding conceptualising and operationalising TC factors. In order to attain more prevailing understating and powerful interpretations on how L2 performance as mediated by TC promotes language learning, a more systematic approach is required to operationalise and investigate the constructs of TC and their interaction with other variables. It is hoped that this study has succeeded in addressing the research gaps reported earlier. It is also hoped that the study has offered novel contributions that can advance current research within TBLT and SLA and has provided guidance for future studies. Finally, it is hoped that doing this research under the umbrella of TBLT will attract more attention to this approach, and advance understanding of TBLT as a sufficient approach to L2 learning and teaching to replace the traditional approaches which are still dominant in some parts of the world including my country Jordan.

List of references

- Ahmadian, M. (2013). Working memory and task repetition in second language oral production. *Asian Journal of English Language Teaching*, 23(1), 37-55.
- Al-Jamal, D., & Al-Jamal, G. (2013). An investigation of the difficulties faced by EFL undergraduates in speaking skills. *English Language Teaching*, 7(1), 19-27.
- Albert, Á. (2011). When individual differences come into play. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 239-265). Amsterdam: John Benjamins.
- Allan, D. (2004). *Oxford Placement Test. University of Cambridge Local Examination Syndicate*. Oxford: Oxford University Press.
- Anderson, J. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, 51(4), 355-365.
- Astington, J., & Baird, J. (2005). *Why language matters for theory of mind*. Oxford: Oxford University Press.
- Atkinson, R., & Shiffrin, R. (1968). Human memory: A proposed system and its control processes. *Psychology of learning and motivation*, 2, 89-195.
- Audacity, T. (2012). Audacity. Audio editor and recorder (Version 2.0). Retrieved from <http://www.audacityteam.org/>
- Baddeley, A. (1986). *Working memory*. Oxford: Oxford University Press.
- Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11), 417-423.
- Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews. Neuroscience*, 4(10), 829-839.
- Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual review of psychology*, 63, 1-29.
- Baddeley, A. (2015). Working memory in second language learning. In Z. Wen, M. Mota & A. McNeill (Eds.), *Working Memory in Second Language Acquisition and Processing* (pp. 17-28). Bristol: Multilingual Matters.
- Baddeley, A., & Hitch, G. (1974). Working memory. *Psychology of learning and motivation*, 8, 47-89.
- Baralt, M. (2013). The impact of cognitive complexity on feedback efficacy during online versus face-to-face interactive tasks. *Studies in second language acquisition*, 35(04), 689-725.
- Baralt, M. (2015). Working memory capacity, cognitive complexity and L2 recasts in online language teaching. In Z. Wen, M. Mota & A. McNeill (Eds.), *Working Memory in*

- Second Language Acquisition and Processing* (pp. 248-269). Bristol: Multilingual Matters.
- Beneš, L., & Jiránek, V. (1976). Pat & Mat, Czech stop-motion animated series. Retrieved from <http://en.patmat.cz/home-pat-and-mat/>
- Boersma, P., & Weenink, D. (2008). Doing phonetics by computer: Praat: ver 4.5.01 [Computer program]. *Computer software*, downloaded from <http://www.fon.hum.uva.nl/praat/>.
- Bonner, S. (1994). A model of the effects of audit task complexity. *Accounting, Organizations and Society*, 19(3), 213-234.
- Botwinick, J., & Storandt, M. (1974). *Memory, related functions and age*. Oxford: Charles C Thomas.
- Braarud, P. Ø., & Kirwan, B. (2010). Task complexity: what challenges the crew and how do they cope. In A. Skjerve & A. Bye (Eds.), *Simulator-based Human Factors Studies Across 25 Years* (pp. 233-251). London: Springer.
- Bratman, M. (1987). *Intention, Plans, and Practical Reasoning*. Cambridge: Harvard University Press.
- Breck, E. (1998). SoundScriber. Michigan: University of Michigan. Retrieved from <http://www-personal.umich.edu/~ebreck/code/sscriber/>
- Breen, M. (1987). Learner contributions to task design. In C. Candlin & D. Murphy (Eds.), *Language learning tasks* (pp. 23-46). London: Prentice Hall.
- Brumfit, C. (1984). *Communicative methodology in language teaching: The roles of fluency and accuracy* (Vol. 129). Cambridge: Cambridge University Press.
- Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (Vol. 32, pp. 21-46). Amsterdam: John Benjamins.
- Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23-48). London: Longman.
- Bygate, M. (2015). *Domains and Directions in the Development of TBLT* (Vol. 8). Amsterdam: John Benjamins.
- Campbell, D. J. (1988). Task complexity: A review and analysis. *Academy of management review*, 13(1), 40-52.
- Candlin, C. (1987). Towards task-based language learning. In C. Candlin & D. Murphy (Eds.), *Language learning tasks* (pp. 5-22). London: Prentice Hall.
- Carroll, J. (1968). The psychology of language testing. In A. Davies (Ed.), *Language testing symposium: A psycholinguistic approach* (pp. 46-69). London: Oxford University Press.

- Case, R., Kurland, M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of experimental child psychology*, 33(3), 386-404.
- CEFR. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Council of Europe. Cambridge, UK: Cambridge University Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed)*. New Jersey: Lawrence Erlbaum Associates.
- Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review*, 12(5), 769-786.
- Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive psychology*, 51(1), 42-100.
- Creswell, J. W. (2015). *A concise introduction to mixed methods research*. London: Sage Publications.
- Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, 19(4), 450-466.
- De Jong, N., & Bosker, H. (2013). *Choosing a threshold for silent pauses to measure second language fluency*. Paper presented at the 6th Workshop on Disfluency in Spontaneous Speech (DiSS).
- De Jong, N., & Vercellotti, M. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, 20(3), 387-404.
- Declerck, M., & Kormos, J. (2012). The effect of dual task demands and proficiency on second language speech production. *Bilingualism: Language and cognition*, 15(04), 782-796.
- Deyzac, E., Logie, R., & Denis, M. (2006). Visuospatial working memory and the processing of spatial descriptions. *British Journal of Psychology*, 97(2), 217-243.
- Dörnyei, Z. (2006). Individual differences in second language acquisition. *AILA Review*, 19(1), 42-68.
- Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press
- Doughty, C., & Long, M. (2000). Eliciting second language speech data. In L. Menn & N. Ratner (Eds.), *Methods of studying language production* (pp. 149-177). London: Lawrence Erlbaum.
- Duez, D. (1985). Perception of silent pauses in continuous speech. *Language and speech*, 28(4), 377-389.
- Elder, C., & Ellis, R. (2009). Implicit and explicit knowledge of an L2 and language proficiency. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.), *Implicit and*

- explicit knowledge in second language learning, testing and teaching* (pp. 167-193). Bristol: Multilingual Matters.
- Ellis, N. C., & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics—Introduction to the special issue. *Applied Linguistics*, 27(4), 558-589.
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.
- Ellis, R. (2005). *Planning and task performance in a second language* (Vol. 11). Amsterdam: John Benjamins Publishing.
- Ellis, R. (2009a). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, 30/4, 474–509.
- Ellis, R. (2009b). Implicit and explicit learning, knowledge and instruction. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (Vol. 42, pp. 3-25). Bristol: Multilingual Matters.
- Ellis, R. (2015). *Understanding Second Language Acquisition* (2nd ed.). Oxford: Oxford university press.
- Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.
- Elyas, T., & Picard, M. (2010). Saudi Arabian educational history: Impacts on English language teaching. *Education, Business and Society: Contemporary Middle Eastern Issues*, 3(2), 136-145.
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General*, 128(3), 309-331.
- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological review*, 102(2), 211-245.
- Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27(3), 464-491.
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: Sage.
- Finardi, K., & Weissheimer, J. (2009). On the Relationship between working memory capacity and L2 speech development. *Signótica*, 20(2), 367-391.
- Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in second language acquisition*, 18(03), 299-323.
- Foster, P., & Skehan, P. (2013). Anticipating a Post-task Activity: The Effects on Accuracy, Complexity, and Fluency of Second Language Performance. *Canadian modern language review*, 69(3), 249-273.

- Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning*, 59(4), 866-896.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354-375.
- Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, 36, 98-116.
- Freed, B. (1995). *Second language acquisition in a study abroad context* (Vol. 9). Amsterdam: John Benjamins Publishing.
- Freed, B. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggensbach (Ed.), *Perspectives on fluency* (pp. 243-265). Michigan: The University of Michigan Press.
- Gaillard, S., & Tremblay, A. (2016). Linguistic Proficiency Assessment in Second Language Acquisition Research: The Elicited Imitation Task. *Language Learning*, 66(2), 419-447.
- Gathercole, S. (1999). Cognitive approaches to the development of short-term memory. *Trends in cognitive sciences*, 3(11), 410-419.
- Geranpayeh, A. (2003). A quick review of the English Quick Placement Test (pp. 8-10). University of Cambridge ESOL.
- Gilabert, R. (2005). *Task complexity and L2 narrative oral production*. Unpublished PhD dissertation. University of Barcelona. Spain.
- Gilabert, R. (2007a). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics in Language Teaching IRAL*, 45(3), 215-240.
- Gilabert, R. (2007b). The simultaneous manipulation of task complexity along planning time and [+/-Here-and-Now]: Effects on L2 oral production. In M. d. p. Mayo (Ed.), *Investigating tasks in formal language learning* (Vol. 20, pp. 44-68). Bristol: Multilingual Matters.
- Gilabert, R., & Barón, J. (2013). The impact of increasing task complexity on L2 pragmatic moves. In A. Mackey & K. McDonough (Eds.), *Second language interaction in diverse educational settings* (Vol. 34, pp. 45-69). Amsterdam: John Benjamins.
- Gilabert, R., Barón, J., & Levkina, M. (2011). Manipulating task complexity across task types and modes. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 105-140). Amsterdam: John Benjamins.
- Gilabert, R., & Muñoz, C. (2010). Differences in attainment and performance in a foreign language: The role of working memory capacity. *International Journal of English Studies*, 10(1), 19-42.
- Gilhooly, K. (2004). Working memory and reasoning. In R. Sternberg & J. Leighton (Eds.), *The nature of reasoning* (pp. 49-77). Cambridge: Cambridge University Press.

- Givón, T. (1998). The functional approach to grammar. In M. Tomasello (Ed.), *The new psychology of language: Cognitive and functional approaches to language structure* (Vol. 1, pp. 41-66). Mahwah, NJ: Erlbaum.
- GoldWave, I. (2009). GoldWave Software (Version V5.70). Retrieved from <http://www.goldwave.com>
- Goo, J. (2010). Working memory and reactivity. *Language Learning*, 60(4), 712-752.
- Graesser, A., McNamara, D., & Louwrese, M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. Sweet & C. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York: Guilford Publications.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30 (4), 461-473.
- Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (Vol. 32, pp. 1-20). Amsterdam: John Benjamins.
- Howatt, A. (1984). *A history of English language teaching*. Oxford: Oxford University Press.
- Hulstijn, J. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and cognition*, 15(2), 422-433.
- Hulstijn, J. (2015). *Language Proficiency in Native and Non-native Speakers: Theory and research*. Amestedam: Benjamins Publishing.
- Hunt, K. W. (1970). Syntactic maturity in schoolchildren and adults. *Monographs of the society for research in child development*, 35(1), 1-67.
- IATBLT. (2015). International Association for Task-Based Language Teaching (IATBLT). from <http://www.tbtl.org/start/>
- Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *The Language Learning Journal*, 44(4), 487-505.
- Ishikawa, T. (2006). The effect of task complexity and language proficiency on task-based language performance. *The Journal of AsiaTEFL*, 3(4), 193-225.
- Ishikawa, T. (2008). The effect of task demands of intentional reasoning on L2 speech performance. *The Journal of Asia TEFL*, 5(1), 29-63.
- Ishikawa, T. (2011). Examining the influence of intentional reasoning demands on learner perceptions of task difficulty and L2 monologic speech. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 307-330). Amsterdam: John Benjamins.

- Iwashita, N. (2001). The effect of learner proficiency on interactional moves and modified output in nonnative–nonnative interaction in Japanese as a foreign language. *System*, 29(2), 267-287.
- Jackson, D., & Suethanapornkul, S. (2013). The Cognition Hypothesis: A Synthesis and Meta-Analysis of Research on Second Language Task Complexity. *Language Learning*, 63(2), 330-367.
- Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63(s1), 87-106.
- Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, 64(4), 809-854.
- Kormos, J. (2006). *Speech production and second language learning*. Mahwah, NJ: Lawrence Erlbaum.
- Kormos, J. (2011). Speech production and the Cognition Hypothesis. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 39-60). Amsterdam: John Benjamins.
- Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145-164.
- Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and cognition*, 11(2), 261-271.
- Kormos, J., & Trebits, A. (2011). Working memory capacity and narrative task performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 267-285). Amsterdam: John Benjamins.
- Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning*, 62(2), 439-472.
- Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, 73(4), 440-464.
- Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, 17(1), 48-60.
- Kuiken, F., & Vedder, I. (2011). Task complexity and linguistic performance in L2 writing and speaking. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 91-104). Amsterdam: John Benjamins.
- Kyllonen, P., & Christal, R. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, 14(4), 389-433.

- Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics*, 607–614.
- Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 579-589.
- Lee, M., & Tedder, M. (2003). The effects of three different computer texts on readers' recall: based on working memory capacity. *Computers in Human Behavior*, 19(6), 767-783.
- Leighton, J. (2004). Defining and describing reason. In R. Sternberg & J. Leighton (Eds.), *The nature of reasoning* (pp. 3-11). Cambridge: Cambridge University Press.
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387-417.
- Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25-42). Michigan: The University of Michigan Press.
- Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge MA: MIT Press.
- Levkina, M., & Gilabert, R. (2012). The effects of cognitive task complexity on L2 oral production. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency investigating complexity, accuracy, and fluency in SLA* (pp. 171-198). Amsterdam: John Benjamins.
- Litosseliti, L. (2010). *Research methods in linguistics*. London: Continuum.
- Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics*, 42(6), 553-568.
- Long, M. (1985). Input and second language acquisition theory. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 377-393). Rowley, MA: Newbury House.
- Long, M. (1998). Focus on form in task-based language teaching. In R. Lambert & E. Shohamy (Eds.), *Language policy and pedagogy* (pp. 179-192). Amsterdam: John Benjamins.
- Long, M. (2015a). Building the road as we travel. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT* (Vol. 8, pp. 1-26). Amsterdam: John Benjamins Publishing.
- Long, M. (2015b). *Second language acquisition and task-based language teaching*. Chichester, West Sussex: Wiley Blackwell.
- Long, M., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly*, 27-56.
- Long, M., & Crookes, G. (1993). Units of analysis in syllabus design: The case for task. In G. Crookes & S. Gass (Eds.), *Task and language learning: Integrating theory and practice* (pp. 9-54). Clevedon: Multilingual Matters.
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2), 190-208.

- Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning*, 60(3), 501-533.
- Malicka, A. (2014). The role of task sequencing in monologic oral production. In P. Robinson (Ed.), *Task sequencing and instructed second language learning* (pp. 71-93). London: Bloomsbury.
- Malicka, A., & Levkina, M. (2012). Measuring task complexity: does L2 proficiency matter. In A. Shehadeh & C. Coombe (Eds.), *Task-based Language Teaching in Foreign Language Contexts: Research and Implementation* (pp. 43-66). Amsterdam: John Benjamins.
- Malicka, A., & Sasayama, S. (2017). *The importance of learning from accumulated knowledge: Findings from a research synthesis on task complexity*. Paper presented at the 7th International Task-Based Language Teaching Conference, Barcelona.
- Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language testing*, 19(1), 85-104.
- McCarthy, P. M., & Jarvis, S. (2010). MTLT, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2), 381-392.
- Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, 16(3), 5-19.
- Meara, P., & Miralpeix, I. (2016). *Tools for Researching Vocabulary*. Bristol, UK: Multilingual Matters.
- Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in second language acquisition*, 20(01), 83-108.
- Merrill, M. (2006). Hypothesized performance on complex tasks as a function of scaled instructional strategies. In J. Elen & R. Clark (Eds.), *Handling complexity in learning environments: Research and theory* (pp. 265-282). Oxford: Elsevier.
- Michel, M. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 141-173). Amsterdam: John Benjamins.
- Mitchell, A. E., Jarvis, S., O'Malley, M., & Konstantinova, I. (2015). Working memory measures and L2 proficiency. In Z. Wen, M. Borges & A. McNeill (Eds.), *Working Memory in Second Language Acquisition and Processing* (Vol. 87, pp. 270-283). Bristol: Multilingual Matters.
- Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. New York: Cambridge University Press.
- Morra, S. (1994). Issues in working memory measurement: Testing for M capacity. *International Journal of Behavioral Development*, 17(1), 143-159.

- Mota, M. (2003). Working memory capacity and fluency, accuracy, complexity, and lexical density in L2 speech production. *Fragmentos*, 24, 69-104.
- Norris, J. (2015). Thinking and acting programmatically in task-based language teaching. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT* (Vol. 8, pp. 27-57). Amsterdam: John Benjamins.
- Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555-578.
- Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press.
- Nunan, D. (2006). *Task-based language teaching*. Ernst: Klett Sprachen.
- Okura, E., & Lonsdale, D. (2012). Working memory's meager involvement in sentence repetition tests. *Cognitive Science Society*, 34, 2132-2137.
- Oller, J. W. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die neueren sprachen*, 75(2), 165-174.
- Pallant, J. (2013). *SPSS survival manual*. London: McGraw-Hill Education.
- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590-601.
- Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, 31(1), 117-134.
- Pang, F., & Skehan, P. (2014). Self-reported planning behaviour and second language performance in narrative retelling. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (Vol. 5, pp. 95-127). Amsterdam: John Benjamins.
- Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in second language acquisition*, 11(01), 63-90.
- Plonsky, L., & Oswald, F. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878-912.
- Prabhu, N. (1987). *Second language pedagogy* (Vol. 20). Oxford: University Press Oxford.
- Préfontaine, Y. (2013). Perceptions of French fluency in second language speech production. *Canadian modern language review*, 69(3), 324-348.
- Purpura, J. (2004). *Assessing grammar* (Cambridge Language Assessment Series). Cambridge: Cambridge University Press.
- Qian, L. (2014). Get it right in the end. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (Vol. 5, pp. 129-154). Amsterdam: John Benjamins.

- Rahman, M., & Alhaisoni, E. (2013). Teaching English in Saudi Arabia: prospects and challenges. *Academic Research International*, 4(1), 112-118.
- Rasinger, S. M. (2013). *Quantitative research in linguistics: An introduction*. London: Bloomsbury.
- Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in second language acquisition*, 31(03), 437-470.
- Révész, A. (2011). Task complexity, focus on L2 constructions, and individual differences: A classroom - based study. *The Modern Language Journal*, 95(1), 162-181.
- Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgements: A Validation Study. *Studies in second language acquisition*, 38(4), 703-737.
- Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of child language*, 14(02), 201-209.
- Richards, J. (1984). The secret life of methods. *TESOL Quarterly*, 18(1), 7-23.
- Richards, J., & Rodgers, T. (2014). *Approaches and methods in language teaching* (3rd ed.). Cambridge: Cambridge University Press.
- Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning*, 45(1), 99-140.
- Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22(1), 27-57.
- Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *IRAL-International Review of Applied Linguistics in Language Teaching*, 43(1), 1-32.
- Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *IRAL-International Review of Applied Linguistics in Language Teaching*, 45(3), 193-213.
- Robinson, P. (2010). Situating and distributing cognition. In M. Pütz & L. Scola (Eds.), *Cognitive processing in second language acquisition: Inside the learner's mind* (Vol. 13, pp. 243-268). Amsterdam: John Benjamins Publishing.
- Robinson, P. (2011a). *Second language task complexity: researching the cognition hypothesis of language learning and performance* (Vol. 2). Amsterdam: John Benjamins.
- Robinson, P. (2011b). Task - based language learning: A review of issues. *Language Learning*, 61(1), 1-36.
- Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and*

- Directions in the Development of TBLT* (Vol. 8, pp. 87-121). Amsterdam: John Benjamins.
- Sagarra, N. (2008). Working memory and L2 processing of redundant grammatical forms. In Z. Han (Ed.), *Understanding second language process* (pp. 133-147). Bristol: Multilingual Matters.
- Salthouse, T. (1992). Working-memory mediation of adult age differences in integrative reasoning. *Memory & Cognition*, 20(4), 413-423.
- Samuda, V., & Bygate, M. (2008). *Tasks in second language learning*. Basingstoke: Palgrave Macmillan.
- Sangarun, J. (2005). The effects of focusing on meaning and form in strategic planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 111-141). Philadelphia: John Benjamins.
- Sauro, J., & Lewis, J. (2016). *Quantifying the user experience: Practical statistics for user research*. Amsterdam: Morgan Kaufmann.
- Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review*, 11, 237-326.
- Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.
- Shah, S., Hussain, M., & Nasseef, O. (2013). Factors Impacting EFL Teaching: An Exploratory Study in the Saudi Arabian Context. *Arab World English Journal*, 4(3), 104-123.
- Shehadeh, A. (2005). Task-based language learning and teaching: Theories and applications. In C. Edwards & J. Willis (Eds.), *Teachers exploring tasks in English language teaching* (pp. 13-30). Hampshire: Palgrave Macmillan.
- Shehdeh, F. (2010). Challenges of teaching English in the Arab world: Why can't EFL programs deliver as expected? *Procedia-Social and Behavioral Sciences*, 2(2), 3600-3604.
- Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, 17(1), 38-62.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
- Skehan, P. (2001). Tasks and language performance. In M. Bygate, M. Swain & P. Skehan (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 167-186). New York: Routledge.
- Skehan, P. (2003). Task-based instruction. *Language teaching*, 36(01), 1-14.
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510-532.

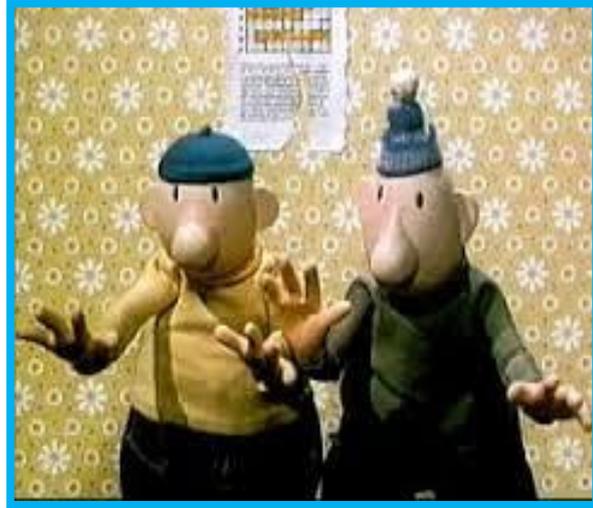
- Skehan, P. (2014). The context for researching a processing perspective on task performance. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (Vol. 5, pp. 1-26). Amsterdam: John Benjamins.
- Skehan, P. (2015a). Limited Attention Capacity and Cognition. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT* (Vol. 8, pp. 123-155). Amsterdam: John Benjamins Publishing.
- Skehan, P. (2015b). Working memory and second language performance: A commentary In Z. Wang, M. Mailce & M. Arthur (Eds.), *Working Memory in Second Language Acquisition and Processing* (Vol. 87, pp. 189-201). Bristol: Multilingual Matters.
- Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1), 93-120.
- Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 183-205). Cambridge: Cambridge University Press.
- Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (Vol. 11, pp. 193-216). Amsterdam: John Benjamins.
- Smith, L., & Abouammoh, A. (2013). Higher education in Saudi Arabia: Reforms, challenges and priorities. In L. Smith & A. Abouammoh (Eds.), *Higher Education in Saudi Arabia. Higher Education Dynamics* (Vol. 40, pp. 1-12). Dordrecht: Springer.
- Swain, M. (1995). Three functions of output in second language learning. *Principle and practice in applied linguistics: Studies in honour of HG Widdowson*, 2(3), 125-144.
- Swales, J. (2009). The concept of task. In K. Van den Branden, M. Bygate & J. Norris (Eds.), *Task-based language teaching* (pp. 41-55). John Benjamins: Amsterdam.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson Education.
- Tauroza, S., & Allison, D. (1990). Speech rates in British English. *Applied Linguistics*, 11(1), 90-105.
- Tavakoli, P. (2009a). Assessing L2 task performance: Understanding effects of task design. *System*, 37(3), 482-495.
- Tavakoli, P. (2009b). Investigating task difficulty: learners' and teachers' perceptions. *International Journal of Applied Linguistics*, 19(1), 1-25.
- Tavakoli, P. (2011). Pausing patterns: differences between L2 learners and native speakers. *ELT journal*, 65(1), 71-79.
- Tavakoli, P. (2014). Storyline complexity and syntactic complexity in writing and speaking tasks. In H. Byrnes & R. Manchón (Eds.), *Task-Based Language Learning – Insights from and for L2 Writing* (Vol. 7, pp. 217-236). Amsterdam: Benjamins Publishing.

- Tavakoli, P. (2016). Speech fluency in monologic and dialogic task performance. *IRAL*, 54(2), 133-151.
- Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, 58(2), 439-473.
- Tavakoli, P., & Foster, P. (2011). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, 61(1), 37-72.
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (Vol. 11, pp. 239-273). Amsterdam: John Benjamins.
- Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research. *Studies in second language acquisition*, 33(3), 339-372.
- Turner, M., & Engle, R. (1989). Is working memory capacity task dependent? *Journal of memory and language*, 28(2), 127-154.
- Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory*, 17(6), 635-654.
- Van den Branden, K. (2006). Introduction: Task-based Language Teaching in a Nutshell. In K. Branden Van den (Ed.), *Task-Based Language Education—From Theory to Practice* (pp. 1-16). Cambridge Cambridge University Press
- Van den Branden, K., Bygate, M., & Norris, J. (2009). *Task-based language teaching: Introducing the reader* (Vol. 1). Amsterdam: John Benjamins
- VanPatten, B. (1990). Attending to form and content in the input. *Studies in second language acquisition*, 12(03), 287-301.
- Wang, Z., & Skehan, P. (2014). Structure, lexis, and time perspective. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (Vol. 5, pp. 155-185). Amsterdam: John Benjamins.
- Wen, Z. (2012). Working memory and second language learning. *International Journal of Applied Linguistics*, 22(1), 1-22.
- Wen, Z. (2015). Working memory in second language acquisition and processing: The phonological/executive model. In Z. Wen, M. Mota & A. McNeill (Eds.), *Working Memory in Second Language Acquisition and Processing* (pp. 41-62). Bristol: Multilingual Matters.
- Wen, Z., Mota, M., & McNeill, A. (2015). *Working memory in second language acquisition and processing* (Vol. 87). Bristol: Multilingual Matters.
- Willis, J. (1996). *A framework for task-based learning*. Harlow: Longman.
- Witton-Davies, G. (2014). *The study of fluency and its development in monologue and dialogue*. Language testing. Unpublished doctoral dissertation. Lancaster University. Lancaster.

- Wood, R. (1986). Task complexity: Definition of the construct. *Organizational behavior and human decision processes*, 37(1), 60-82.
- Wright, C. (2010). *Role of Working Memory in SLA*. Saarbrücken: VDM Publishing House.
- Wright, C. (2015). Working Memory and L2 Development Across the Lifespan: A Commentary. In Z. Wen, M. Mota & A. McNeill (Eds.), *Working Memory in Second Language Acquisition and Processing* (Vol. 87, pp. 285-298). Bristol: Multilingual Matters.
- Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, 46(4), 680-704.
- Yuan, F., & Ellis, R. (2003). The effects of pre - task Planning and on - Line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 24(1), 1-27.

Appendix 1: Task instructions (English version)

Task Instructions



Task 1: Tell and Describe

You are going to watch a two-minute video clip from Pat & Mat series. As you are watching, I need you to tell and describe what is happening in the clip in details in English.

Imagine that you are telling what is happening to someone who is not watching the video. You need to try to speak while the video runs. When the video stops, you have 10 seconds to finish your speech. You are going to use a headphone to record your voice.

Note that Pat is wearing a yellow sweater, whereas Mat is wearing a green sweater.

Tasks Instructions



Task 2: Explain intentional reasoning & its results

You are going to watch a two-minute video clip from Pat & Mat series. As you are watching, I need you to tell and describe what is happening in the clip in details in English.

Also, I need you to explain why Pat & Mat are solving their problems or behaving in certain ways. Explain their intentions and the unpredicted results or consequences of their actions.

Imagine that you are telling what is happening and why it is happening to someone who is not watching the video. You need to try to speak while the video runs. When the video stops, you have 10 seconds to finish your speech. You are going to use a microphone to record your voice.

Note that Pat is wearing a yellow sweater, whereas Mat is wearing a red sweater.

Appendix 2: Task instructions (Arabic version)

تعليمات المهمة الأولى



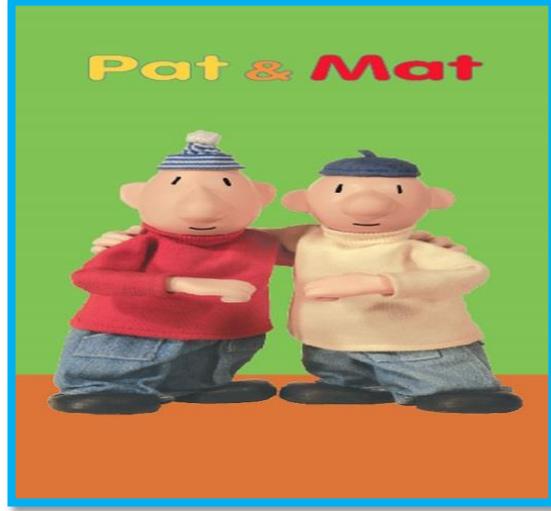
المهمة الأولى: سرد الأحداث ووصفها

سوف تشاهد مقطع فيديو مدته تقريبا 120 ثانية من مسلسل زينغو ورينغو (بات و مات).
أثناء مشاهدة الفيديو أريد منك أن تخبرني وتصف ما يحدث في المقطع بالتفصيل
وباللغة الإنجليزية.

تخيل أنك تسرد ما يحدث لشخص لا يشاهد الفيديو. يجب عليك أن تستمر في التحدث أثناء
تشغيل الفيديو. عندما يتوقف الفيديو لديك 10 ثوان لتنتهي كلامك. سوف تستخدم
مايكروفون لتسجيل صوتك.

ملاحظة: بات يرتدي بلوزة صفراء ومات يرتدي بلوزة خضراء.

تعليمات المهمة الثانية



المهمة الثانية: شرح النوايا وقراءة الأفكار

سوف تشاهد مقطع فيديو مدته تقريبا 120 ثانية من مسلسل زينغو وريغو (بات و مات).
أثناء مشاهدة الفيديو أريد منك أن تخبرني وتصف ما يحدث في المقطع بالتفصيل
وباللغة الإنجليزية.

أريد منك أيضاً أن تشرح لماذا بات ومات يقومون بحل مشاكلهم أو يتصرفون بطريقة
معينة. عليك أن تحاول أن تكشف عن ما ينوون فعله وما هي النتائج الغير متوقعة أو
العواقب المترتبة على ما يقومون به.

تخيل أنك تسرد ما يحدث ولماذا يحدث لشخص لا يشاهد الفيديو. يجب عليك أن تستمر في
التحدث أثناء تشغيل الفيديو. عندما يتوقف الفيديو لديك 10 ثوان لتنتهي كلامك. سوف
تستخدم مايكروفون لتسجيل صوتك.

ملاحظة: بات يرتدي بلوزة صفراء ومات يرتدي بلوزة حمراء.

Appendix 3: Ethics Forms

School of Literature and Languages

Department of English Language and Applied Linguistics



ETHICS COMMITTEE

Consent Form

Project title: **“The effect of cognitive task complexity manipulated by reasoning demands on second language learners’ oral performance: interaction with language proficiency and working memory capacity”**

I understand the purpose of this research and I understand what is required of me;

I have read and understood the Information Sheet relating to this project, which has been explained to me by *Anas Ahmad Mohammad Awwad*. I agree to the arrangements described in the Information Sheet so far as they relate to my participation.

I understand that my participation is entirely voluntary and that I have the right to withdraw from the project at any time.

I have received a copy of this Consent Form and the Information Sheet.

Name:

Signed:

Date:

Researcher:
Anas Ahmad Mohammad Awwad
Email: a.a.m.awwad@reading.ac.uk
Supervisor:
Dr. Parvaneh Tavakoli
Email: p.tavakoli@reading.ac.uk

Department of English Language and Applied Linguistics
HUMSS Building
The University of Reading
Whiteknights, PO Box 219 Reading RG6 6AW
Phone: +44 (0)118 975 6506
Email: appling@reading.ac.uk

INFORMATION SHEET (Student's copy)

The purpose of this study is to investigate the effect of task complexity on second language speech production by performing different monologic narrative tasks. The participants' speech production will be recorded and analyzed to measure the complexity, accuracy and fluency of their performance.

You have been selected to participate in this study because your second language is English. Your language proficiency will be assessed prior to the study by employing Oxford Placement Test and an elicited imitation task. A backward digit span test will be used to measure your working memory. Then, you will be required to retell two short stories in English based on two video clips at your school and your voice will be recorded. You will be also asked to complete a questionnaire on your perceptions of task difficulty.

The collected data will be securely kept on a password-protected computer or in a locked drawer. Only the researcher, the supervisors and the examiners will have an access to the data. The data will be used for academic purposes only and will be anonymous. The data will be destroyed after the completion of the PhD thesis. Your privacy and confidentiality will be carefully observed and you have the right to withdraw from the study at any time you wish to.

This project has been subject to ethical review by the School Ethics and Research Committee, and has been allowed to proceed under the exceptions procedure as outlined in paragraph 6 of the University's *Notes for Guidance* on research ethics.

If you have any queries or wish to clarify anything about the study, please feel free to contact my supervisor at the address above or by email at p.tavakoli@reading.ac.uk

Signed

Appendix 4: Language Background Questionnaire

Name: _____

Age: _____

Year Group (class): _____

Country: _____

First language: _____

Last term average (English subject): _____

Years lived in an English-speaking country: _____

Years of learning English as a foreign language: _____

Notes: _____

Appendix 5: Task Difficulty Questionnaire

Name:

Learner's perceptions of task difficulty

The aim of this questionnaire is to give your opinion about the level of difficulty of the two tasks you have already performed. Your feedback is very important for this research.

Instructions

Circle the adjective that best describes the level of difficulty of the tasks you have already performed:

Task one: **Tell and describe (The Outdoor Lunch)**

1. **Very easy** 2. **Easy** 3. **Difficult** 4. **Very difficult**

Give reasons for you answer.

.....
.....

Task two: **Explain intentions and thoughts (The Flying Car)**

1. **Very easy** 2. **Easy** 3. **Difficult** 4. **Very difficult**

Give reasons for you answer.

.....
.....

Thank you for your participation

Researcher: **Anas Ahmad Awwad**

Email: **a.a.m.awwad@reading.ac.uk**

Appendix 6: The coding symbols

| = the end of AS-Unit

:: = the end of a clause

errfr = error-free clause

= error (semantic, syntactic, intonational, or nativelike use)

highlight = incorrect pronunciation

“” = false starts

* = repetition

~ = reformulation

∞ = Replacements

Λ = Hesitations

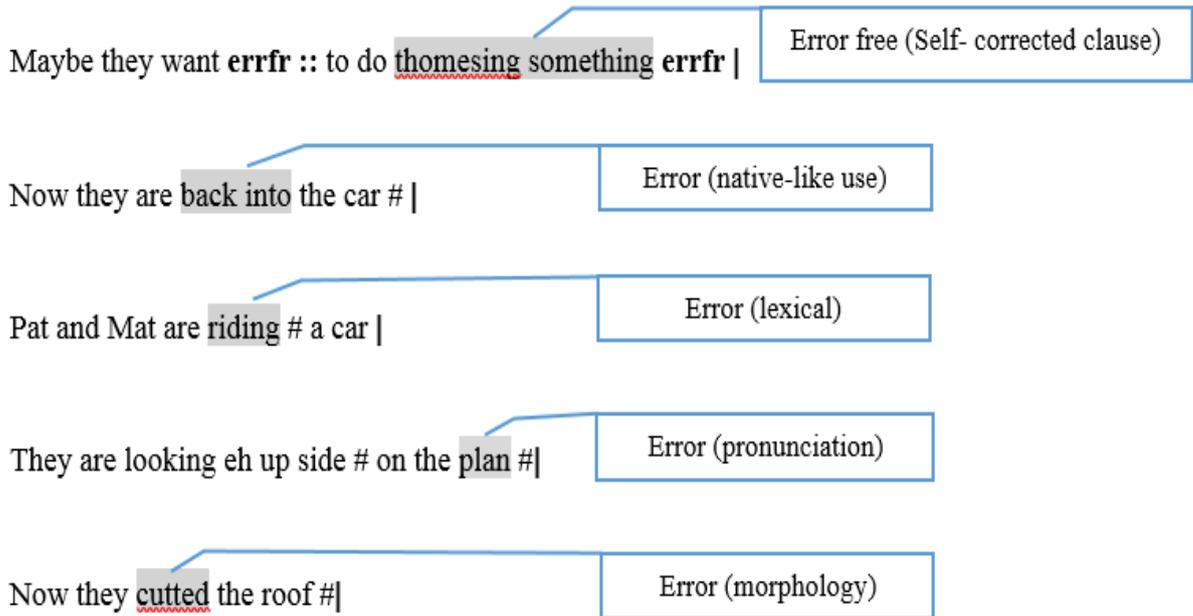
× = inaudible word

⌘ = Abandoned utterance

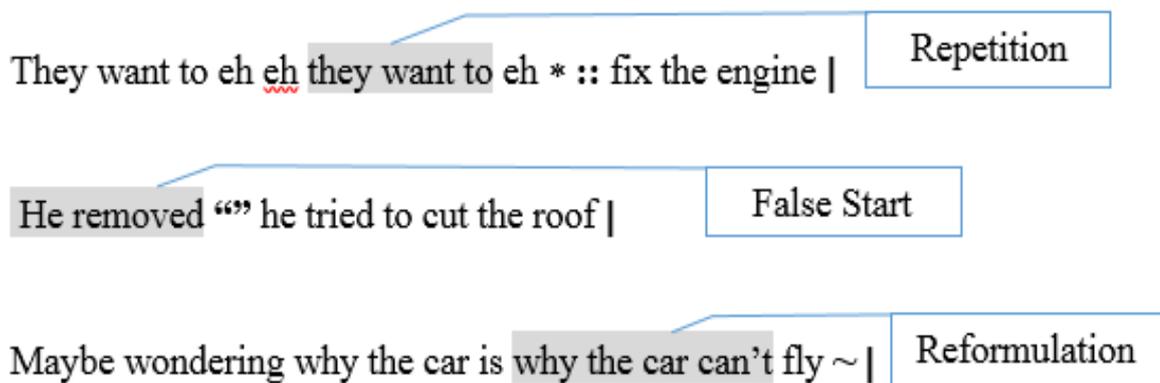
() = Silent pause length

eh = Filled pauses

Appendix 7: Examples of types of errors



Appendix 8: Examples of repairs



Appendix 9: Samples of the coded data (Study One)

| | | | |
|-----------------------------|---------------------------------|--------------------------------------|--------------------------|
| ST: 1 | Task: Rainy lunch (- IR) | Time: 1.38 min | No. of words: 127 |
| Total AS units: 17 | | Total clauses: 25 | |
| Total fillers: 22 | | No. of error-free clauses: 14 | |
| Total repetitions: 1 | | No. of errors: 11 | |
| False start: 1 | | Reformulation: 1 | |

Pat and Mat are eh sitting outside (0.67) **errfr** | eh Pat is knitting **errfr ::** and listening to music **errfr ::** while Mat is relaxing **errfr::** and reading a book (4.13) **errfr** | eh (1.25) Mat eh stands **errfr ::** and looking at the clock (0.80) # | and eh told his friends (0.71) eh the time (1.18) # | There in the kitchen (2.11) # eh they have food eh and a frying pan (4.33) | He is eh (0.46) pouring water in the eh (1.11) cooking pot **errfr ::** putting food in it **errfr** | He's cooking (2.68) **errfr** | He starting a fire (1.95) # | and he is looking for a match (2.10) **errfr** | He started the fire eh (2.55) light off (3.10)# | eh the eh sky started # :: to rain and looking eh up # | eh they bor they bring ~ an umbrella **errfr** | and went inside eh the eh cottage # | eh they're moving the stuff inside **errfr** | He put the chicken in the oven # | And it's "" eh he's waiting **errfr ::** for it to finish # | He is matching # :: eh to light eh the eh the * oven **errfr** |

| | |
|--|-------------------------------------|
| Percentage of error free clauses: 56% | Mean length of AS unit: 7.47 |
| Errors per 100 words: 8.6 | Mean length of clauses: 5.08 |
| Dysfluencies per minute: 1.8 | Ratio of subordination: 1.47 |
| Number of silent mid-clause pauses (> 0.25 sec) / minute: 6 | |
| Mean length of mid-clause pauses / minute: 1.3 sec | |
| Number of silent end-clause pauses (> 0.25 sec) / minute: 8 | |
| Mean length of end-clause pauses / minute: 2.3 sec | |
| Number of filled pauses per minute: 13.5 | |

| | | | |
|-----------------------------|---------------------------------|--------------------------------------|--------------------------|
| ST: 11 | Task: Rainy lunch (- IR) | Time: 1.40 min | No. of words: 153 |
| Total AS units: 21 | | Total clauses: 29 | |
| Total fillers: 19 | | No. of error-free clauses: 12 | |
| Total repetitions: 1 | | No. of errors: 20 | |
| False start: 0 | | Reformulation: 1 | |

eh Pat now is eh knitting **errfr ::** and listen to music (0.41) #| eh Mat is reading a story (0.29) behind their eh cottage (3.68) **errfr** | eh Mat is looking at the clock (0.90) **errfr** | and eh tell eh Pat # :: that he is going to cook (1.49) **errfr** | eh he wants **errfr ::** to cook an eggs # and chicken (0.60) and sausage (0.45) in the eh fire (1.00) # | but they don't have an oven (2.58) **errfr** | He start # :: to put water in the eh (0.45) cooking jar (0.41) eh cooking (0.36) pot (3.21) ~ **errfr** | He start # :: to eh put the wood (0.88) **errfr ::** to fire it (0.95) # | He want a match (1.81) # | He start a fire (2.70) #| It's # doesn't works (2.00) # | eh Mat told (0.41) # :: he is looking eh (0.85) to find a match # | It's raining **errfr** | so he can't start a fire **errfr** | He put # an umbrella in the oven # | but it doesn't work **errfr** | eh he put # it in the oven in the house in his house * | and put # the chicken in the oven | He's waiting right now **errfr** | Pat put # the eh coal in the oven | So they start to eh cooking # :: and eh start to fire # |

| | |
|---|-------------------------------------|
| Percentage of error free clauses: 41.37% | Mean length of AS unit: 7.28 |
| Errors per 100 words: 13.07 | Mean length of clauses: 5.2 |
| Dysfluencies per minute: 1.2 | Ratio of subordination: 1.38 |
| Number of silent mid-clause pauses (> 0.25 sec) / minute: 7 | |
| Mean length of mid-clause pauses / minute: 0.48 sec | |
| Number of silent end-clause pauses (> 0.25 sec) / minute: 13 | |
| Mean length of end-clause pauses / minute: 1.69 sec | |
| Number of filled pauses per minute: 11.4 | |

| | | | |
|-----------------------------|---------------------------------|--------------------------------------|--------------------------|
| ST: 20 | Task: Rainy lunch (- IR) | Time: 1.31 min | No. of words: 171 |
| Total AS units: 24 | | Total clauses: 31 | |
| Total fillers: 13 | | No. of error-free clauses: 12 | |
| Total repetitions: 2 | | No. of errors: 22 | |
| False start: 2 | | Reformulation: 3 | |

Mat and Pat are relaxing outside (0.41) **errfr** | Mat is read eh (0.29) Mat is reading ~ a book **errfr** | and Pat is cook is eh knitting (1.27) ~ **errfr** | He's listening to music (0.56) **errfr** | And they sitting out (0.41) # of the cottage (0.83) | Pat Mat is looking at his clock (0.44) # | It's lunchtime (1.10) **errfr** | Tell # Mat to eh come :: to eh help him to cook (1.33) **errfr** | They're looking at the food **errfr** | They have chicken sausage (0.90) and chicken pot cooking pot (1.91) ~ **errfr** | eh he put # the water (0.33) in a cooking pot (1.10) | And wait (0.82) for eh put # potato (2.00) and eh (0.47) frying frying * pan (0.78) # | He's now putting wood (0.29) **errfr** :: to eh to * fire (0.59) to make it fire (0.99) **errfr** | He have matches (1.41) # :: and want to fire it (0.67) # | But it (1.55) "" there is rain # | so he can't fire the wood (5.23) # | Mat "" eh the rain is falling **errfr** | eh Mat bring # umbrella # :: to make food # warm :: and eh don't # water comes # on it | They took # the food inside | And put it on oven # :: to make it ready to eh eat **errfr** | He put # chicken in the eh oven | and want fire # | So he put # wood inside it | And bring match # :: to fire it # |

| | |
|---|-------------------------------------|
| Percentage of error free clauses: 38.7% | Mean length of AS unit: 7.12 |
| Errors per 100 words: 12.9 | Mean length of clauses: 5.5 |
| Dysfluencies per minute: 4.6 | Ratio of subordination: 1.29 |
| Number of silent mid-clause pauses (> 0.25 sec) / minute: 8 | |
| Mean length of mid-clause pauses / minute: 0.86 sec | |
| Number of silent end-clause pauses (> 0.25 sec) / minute: 15 | |
| Mean length of end-clause pauses / minute: 1.22 sec | |
| Number of filled pauses per minute: 8.5 | |

| | | | |
|-----------------------------|-------------------------------|-------------------------------------|--------------------------|
| ST: 1 | Task: Flying Car (+IR) | Time: 1.30 min | No. of words: 170 |
| Total AS units: 26 | | Total clauses: 36 | |
| Total fillers: 30 | | No of error-free clauses: 26 | |
| Total repetitions: 0 | | No of errors: 11 | |
| False starts: 1 | | Reformulation: 1 | |

Pat and Mat are driving a car (0.41) **errfr** | I think eh **errfr** :: they eh are trying to fly **errfr** | They have a lot of luggage eh in the eh car (1.24) **errfr** | They stopped # | and eh I think eh **errfr** :: there is something wrong **errfr** | They want to eh make sure **errfr** :: that the err car can fly (0.87) **errfr** | They eh want it like an aeroplane (0.29)**errfr** | and (0.41) eh the (1.00) eh “” (1.02) they are err making wings for err (0.86) **errfr** | They’re cutting the eh roof **errfr** | so it’s eh like wings # | so they can fly (0.75) **errfr** | They’re eh start flying # | but I think **errfr** :: eh it seems like **errfr** :: they eh (0.63) will not **errfr** :: because there is a lot of luggage **errfr** :: and it is heavy(1.88) **errfr** | So they eh remove the luggage **errfr** :: and they’re trying to fly **errfr** | It won’t work (0.55) **errfr** | So they wanted to remove the engine (0.38) # | so it eh would be lighter #| so they can fly (1.25) **errfr** | eh They are removing it **errfr** | So eh (0.67) they are trying again **errfr** | but eh the car I think that **errfr** :: it won’t start **errfr** :: and eh because ~ there is no engine **errfr** | So eh eh the truck came by # | and eh they eh eh tow the rope eh with it # | He drove the car a #| and the car started # to flying # | and it worked # |

| | |
|--|-------------------------------------|
| Percentage of error free clauses: 72.22% | Mean length of AS unit: 6.50 |
| Errors per 100 words: 6.4 | Mean length of clauses: 4.7 |
| Dysfluencies per minute: 1.3 | Ratio of subordination: 1.38 |
| Number of silent mid-clause pauses (> 0.25 sec) /minute: 5 | |
| Mean length of mid-clause pauses / minute: 0.75 sec | |
| Number of silent end-clause pauses (> 0.25 sec) /minute: 10 | |
| Mean length of end-clause pauses / minute: 0.85 sec | |
| Number of filled pauses per minute: 20 | |

| | | | |
|-----------------------------|-------------------------------|-------------------------------------|--------------------------|
| ST: 11 | Task: Flying Car (+IR) | Time: 1.37 min | No. of words: 164 |
| Total AS units: 23 | | Total clauses: 30 | |
| Total fillers: 18 | | No of error-free clauses: 17 | |
| Total repetitions: 1 | | No of errors: 13 | |
| False start: 1 | | Reformulation: 0 | |

Eh Pat and Mat are eh driving eh an orange car (0.44) **errfr** | They are trying to eh fly (0.45) **errfr** | The car have # a roof and eh a luggage (2.00) # | They are stopped (0.73) # | Maybe there is “” eh (0.75) the eh (1.58) the * engine start off (1.60) # | They saw eh (0.99) a rocket or eh a (0.29) plane (3.93) # | They cut their eh (1.55) they ~ wing # | Maybe they want to fly (1.66) **errfr** | Oh yes it start to fly # | But maybe it’s too heavy to fly (3.14) **errfr** | They put their luggage out of the car **errfr** (0.65) :: because to make eh it lighter (1.78) # | It stopped again (0.60) # | May be the engine doesn’t work (0.85) **errfr** :: or it’s heavy (1.21) **errfr** | I think :: it’s heavy (1.61) **errfr** | They start to fix it **errfr** | No eh they put it out # | Now they can’t start eh the car **errfr** :: because there’s no engine **errfr** | There’s a truck **errfr** :: to eh tow the car **errfr** | They put the rope **errfr** | Maybe they want **errfr** :: to eh pull the car **errfr** | It seems like it’s fly # | Yes it’s fly # :: but there is a rope between the truck and eh the car **errfr** | eh maybe they want to fly like eh airplane # |

| | |
|---|-------------------------------------|
| Percentage of error free clauses: 56.66% | Mean length of AS unit: 7.1 |
| Errors per 100 words: 7.9 | Mean length of clauses: 5.46 |
| Dysfluencies per minute: 1.23 | Ratio of subordination: 1.30 |
| Number of silent mid-clause pauses (> 0.25 sec) / minute: 5 | |
| Mean length of mid-clause pauses / minute: 1.03 sec | |
| Number of silent end-clause pauses (> 0.25 sec) / minute: 14 | |
| Mean length of end-clause pauses / minute: 1.47 sec | |
| Number of filled pauses per minute: 11 | |

| | | | |
|-----------------------------|-------------------------------|--------------------------------------|--------------------------|
| ST: 20 | Task: Flying Car (+IR) | Time: 1.28 min | No. of words: 196 |
| Total AS units: 21 | | Total clauses: 34 | |
| Total fillers: 16 | | No. of error-free clauses: 23 | |
| Total repetitions: 1 | | No. of errors: 12 | |
| False start: 0 | | Reformulation: 5 | |

Another day with Pat and Mat (0.72) eh today they want to go outside eh a trip (0.55) # | They taking (0.53) # a luggage # with them (0.40) | And they have a roof in the car (0.61) **errfr** | They are moving (1.29) **errfr** | And (0.36) may Maybe ~ they want (0.52) **errfr ::** to go another place to trip or to eh relax (0.70) # | They are watching up (0.98) # | The eh they look at plane # | And they think **errfr ::** that they can fly like it (2.93) **errfr** | They (0.74) they * cut the roof **errfr ::** to make it eh (0.99) wings (0.38) to fly (0.53) **errfr** | And eh (0.68) they go (0.46) **errfr** | But the car can't **errfr ::** because it's too heavy with the luggage (0.77) **errfr** | So I think **errfr ::** they will (0.32) throw the luggage out (1.17) **errfr** | They are taking the luggage (0.58) from the car **errfr ::** to make it eh eh light (2.69) **errfr** | Eh oh they think that **errfr ::** the eh the heavy eh the engine is too heavy ~ **errfr** | So they went they want ~ **errfr ::** to take it off (0.55) # | But the car can't move with eh without engine (1.54) ~ **errfr** | What will what will they do **errfr** | Hhh the car won't move **errfr ::** because no more engine # | Now eh they bring a truck # :: to tow them with eh the the eh eh put a rope on it ~ # | And tell the truck # :: to move fast to fly **errfr** | The car fly # :: because it's too light **errfr ::** and the truck is moving fast **errfr** |

| | |
|---|------------------------------------|
| Percentage of error free clauses: 67.6% | Mean length of AS unit: 9.3 |
| Errors per 100 words: 6.1 | Mean length of clauses: 5.7 |
| Dysfluencies per minute: 4 | Ratio of subordination: 1.6 |
| Number of silent mid-clause pauses (> 0.25 sec) / minute: 9 | |
| Mean length of mid-clause pauses / minute: 0.58 sec | |
| Number of silent end-clause pauses (> 0.25 sec) / minute: 15 | |
| Mean length of end-clause pauses / minute: 1.05 sec | |
| Number of filled pauses per minute: 10.9 | |

Appendix 10: Oxford Placement Test



University of Cambridge
Local Examination Syndicate

OXFORD
University Press

Name:

Date:

Quick placement test

Version 2

The test is divided into two parts:

Part 1 (Questions 1- 40)

Part 2 (Questions 41 – 60)

Time: 45 minutes

Quick Placement Test

Part 1

Question 1 – 5

- ❖ Where can you see these notices?
- ❖ For questions **1** to **5**, mark one letter **A,B** or **C** on your **Answer Sheet**.

| | | | | | |
|--|------------------------|------------------------|----------|----------|----------|
| 1. YOU CAN LOOK, BUT DON'T TOUCH THE PICTURES | | | A | B | C |
| A▶ in an office | B▶ in a cinema | C▶ in a museum | | | |
| 2. PLEASE GIVE THE RIGHT MONEY TO THE DRIVER | | | A | B | C |
| A▶ in a bank | B▶ on a bus | C▶ in a cinema | | | |
| 3. NO PARKING PLEASE | | | A | B | C |
| A▶ in a street | B▶ on a book | C▶ on a table | | | |
| 4. CROSS BRIDGE FOR TRAINS TO EDINBURGH | | | A | B | C |
| A▶ in a bank | B▶ in a garage | C▶ in a station | | | |
| 5. KEEP IN A COLD PLACE | | | A | B | C |
| A▶ on clothes | B▶ on furniture | C▶ on food | | | |

Question 6 –10

- ❖ In this section you must choose the word which best fits each space in the text below.
- ❖ For questions **6** to **10**, mark **one** letter **A**, **B**, or **C** on your Answer Sheet

THE STARS

There are millions of stars in the sky. If you look **(6)**.....the sky on a clear night, it is possible to see about 3000 stars. They look small, but they are really **(7)**.....big hot balls of burning gas. Some of them are huge, but others are much smaller, like our planet Earth. The biggest stars are very bright, but they only live for a short time. Every day new stars **(8)**.....born and old stars die. All the stars are very far away. The light from the nearest star takes more **(9)**.....four years to reach Earth. Hundreds of years ago, people **(10)**.....stars, like the North Star, to know which direction to travel in. Today you can still see that star.

| | | | | | |
|-----------------|-----------------|------------------|----------|----------|----------|
| 6. | | | A | B | C |
| A ► at | B ► up | C ► on | | | |
| 7. | | | A | B | C |
| A ► very | B ► too | C ► much | | | |
| 8. | | | A | B | C |
| A ► is | B ► be | C ► are | | | |
| 9. | | | A | B | C |
| A ► that | B ► of | C ► than | | | |
| 10. | | | A | B | C |
| A ► use | B ► used | C ► using | | | |

Question 11 - 15

- ❖ In this section you must choose the word which best fits each space in the texts.
- ❖ For questions 11 to 20, mark one letter A, B, C or D on your Answer Sheet.

Good smiles ahead for young teeth

Older Britons are the worst in Europe when it comes to keeping their teeth. But British youngsters **(11)**.....more to smile about because **(12)**.....teeth are among the best. Almost 80% of Britons over 65 have lost all or some **(13)**.....their teeth according to a World Health Organisation survey. Eating too **(14)**.....sugar is part of the problem. Among **(15)**....., 12-year-olds have on average only three missing, decayed or filled teeth.

| | | | | | | | |
|-------------------|------------------|--------------------|-------------------|----------|----------|----------|----------|
| 11. | | | | A | B | C | D |
| A▶ getting | B▶ got | C▶ have | D▶ having | | | | |
| 12. | | | | A | B | C | D |
| A▶ their | B▶ his | C▶ them | D▶ theirs | | | | |
| 13. | | | | A | B | C | D |
| A▶ from | B▶ of | C▶ among | D▶ between | | | | |
| 14. | | | | A | B | C | D |
| A▶ much | B▶ lot | C▶ many | D▶ deal | | | | |
| 15. | | | | A | B | C | D |
| A▶ person | B▶ people | C▶ children | D▶ family | | | | |

Question 16 - 20

Christopher Columbus and the New World

On August 3, 1492, Christopher Columbus set sail from Spain to find a new route to India, China and Japan. At this time most people thought you would fall off the edge of the world if you sailed too far. Yet sailors such as Columbus had seen how a ship appeared to get lower and lower on the horizon as it sailed away. For Columbus this **(16)**.....that the world was round. He **(17)**.....to his men about the distance travelled each day. He did not want them to think that he did not **(18)**.....exactly where they were going. **(19)**....., on October 12, 1492, Columbus and his men landed on a small island he named San Salvador. Columbus believed he was in Asia, **(20)**.....he was actually in the Caribbean.

| | | | | | | | |
|----------------|---------------------|-------------------|------------------|----------|----------|----------|----------|
| 16. | | | | A | B | C | D |
| A▶ made | B▶ pointed | C▶ was | D▶ proved | | | | |
| 17. | | | | A | B | C | D |
| A▶ lied | B▶ told | C▶ cheated | D▶ asked | | | | |
| 18. | | | | A | B | C | D |
| A▶ find | B▶ know | C▶ think | D▶ expect | | | | |
| 19. | | | | A | B | C | D |
| A▶ Next | B▶ Secoundly | C▶ Finally | D▶ Once | | | | |
| 20. | | | | A | B | C | D |
| A▶ as | B▶ but | C▶ because | D▶ if | | | | |

Question 21 - 30

- ❖ In this section you must choose the word or phrase which best completes each sentence.
- ❖ For questions 21 to 40, mark one letter A, B, C or D on your Answer Sheet.

| | | | | | | | |
|--|---------------------|--------------------|-----------------------|----------|----------|----------|----------|
| 21. The children won't go to sleep.....we leave a light on outside their bedroom. | | | | A | B | C | D |
| A▶ except | B▶ otherwise | C▶ unless | D▶ but | | | | |
| 22. I'll give you my spare keys in case you.....home before me. | | | | A | B | C | D |
| A▶ would get | B▶ got | C▶ will get | D▶ get | | | | |
| 23. My holiday in Paris gave me a great.....to improve my French accent. | | | | A | B | C | D |
| A▶ occasion | B▶ chance | C▶ hope | D▶ possibility | | | | |
| 24. The singer ended the concert.....her most popular song. | | | | A | B | C | D |
| A▶ by | B▶ with | C▶ in | D▶ as | | | | |
| 25. Because it had not rained for several months, there was a.....of water. | | | | A | B | C | D |
| A▶ shortage | B▶ drop | C▶ scare | D▶ waste | | | | |
| 26. I've always.....you as my best friend. | | | | A | B | C | D |
| A▶ regarded | B▶ thought | C▶ meant | D▶ supposed | | | | |
| 27. She came to live her.....a month ago. | | | | A | B | C | D |
| A▶ quite | B▶ beyond | C▶ already | D▶ almost | | | | |
| 28. Don't make such a.....! The dentist is only going to look at your teeth. | | | | A | B | C | D |
| A▶ fuss | B▶ trouble | C▶ worry | D▶ reaction | | | | |
| 29. He spent a long time looking for a tie which.....with his new shirt. | | | | A | B | C | D |
| A▶ fixed | B▶ made | C▶ went | D▶ wore | | | | |
| 30. Fortunately,.....from a bump on the head, she suffered no serious injuries from her fall. | | | | A | B | C | D |
| A▶ other | B▶ except | C▶ besides | D▶ apart | | | | |

Question 31 – 40

| | | | | | | | |
|---|----------------------|-----------------------|---------------------|----------|----------|----------|----------|
| 31. She had changed so much that.....anyone recognised her. | | | | A | B | C | D |
| A▶ almost | B▶ hardly | C▶ not | D▶ nearly | | | | |
| 32.teaching English, she also writes children´s books. | | | | A | B | C | D |
| A▶ Moreover | B▶ As well as | C▶ In addition | D▶ Apart | | | | |
| 33. It was clear that the young couple were.....of taking charge of the restaurant. | | | | A | B | C | D |
| A▶ responsible | B▶ reliable | C▶ capable | D▶ able | | | | |
| 34. The book.....of ten chapters, each one covering a different topic. | | | | A | B | C | D |
| A▶ comprises | B▶ includes | C▶ consists | D▶ contains | | | | |
| 35. Mary was disappointed with her new shirt as the colour.....very quickly. | | | | A | B | C | D |
| A▶ bleached | B▶ died | C▶ vanished | D▶ faded | | | | |
| 36. National leaders from all over the world are expected o attend the.....meeting. | | | | A | B | C | D |
| A▶ peak | B▶ summit | C▶ top | D▶ apex | | | | |
| 37. Jane remained calm when she won the lottery and.....about her business as if nothing had happened. | | | | A | B | C | D |
| A▶ came | B▶ brought | C▶ went | D▶ moved | | | | |
| 38. I suggest we.....outside the stadium tomorrow at 8.30. | | | | A | B | C | D |
| A▶ meeting | B▶ meet | C▶ met | D▶ will meet | | | | |
| 39. My remarks were.....as a joke, but she was offended by them. | | | | A | B | C | D |
| A▶ pretended | B▶ thought | C▶ meant | D▶ supposed | | | | |
| 40. You ought to take up swimming for the.....of your health. | | | | A | B | C | D |
| A▶ concern | B▶ relief | C▶ sake | D▶ cause | | | | |

Part 2

Do not start this part unless told to do so by your test supervisor

Questions 41 – 45

- ❖ In this section you must choose the word which best fits each space in the texts.
- ❖ For questions 41 to 45, mark one letter A, B, C or D on your Answer Sheet.

CLOCKS

The clock was the first complex mechanical machinery to enter the home, **(41)**.....it was too expensive for the **(42)**.....person until the 19th century, when **(43)**.....production techniques lowered the price. Watches were also developed, but they **(44)**.....luxury items until 1868, When the first cheap pocket watch was designed in Switzerland. Watches later became **(45)**.....available, and Switzerland became the world’s leading watch manufacturing centre for the next 100 years.

| | | | | | | | |
|-------------------|--------------------|---------------------|--------------------|----------|----------|----------|----------|
| 41. | | | | A | B | C | D |
| A▶ despite | B▶ although | C▶ otherwise | D▶ average | | | | |
| 42. | | | | A | B | C | D |
| A▶ average | B▶ medium | C▶ general | D▶ common | | | | |
| 43. | | | | A | B | C | D |
| A▶ vast | B▶ large | C▶ wide | D▶ mass | | | | |
| 44. | | | | A | B | C | D |
| A▶ lasted | B▶ endured | C▶ kept | D▶ remained | | | | |
| 45. | | | | A | B | C | D |
| A▶ mostly | B▶ chiefly | C▶ greatly | D▶ widely | | | | |

Questions 46 - 50

Dublin City Walks

What better way of getting to know a new city than by walking around it? Whether you choose the Medieval Walk, which will **(46)**.....you to the 1000 years ago, find out about the more **(47)**.....history of the city on the Eighteenth Century Walk, or meet the ghosts of Dublin’s many writers on The Literary Walk, we know you will enjoy the experience. Dublin City Walks **(48)**.....twice daily. Meet your guide at 10.30 a.m. or 2.30 p.m. at the Tourist Information Office. No advance **(49)**.....is necessary. Special **(50)**.....are available for families, children and parties of more than ten people.

| | | | | | | | |
|----------------------|---------------------|-------------------|--------------------|----------|----------|----------|----------|
| 46. | | | | A | B | C | D |
| A▶ introduce | B▶ present | C▶ move | D▶ show | | | | |
| 47. | | | | A | B | C | D |
| A▶ near | B▶ late | C▶ recent | D▶ close | | | | |
| 48. | | | | A | B | C | D |
| A▶ take place | B▶ occur | C▶ work | D▶ function | | | | |
| 49. | | | | A | B | C | D |
| A▶ paying | B▶ reserving | C▶ warning | D▶ booking | | | | |
| 50. | | | | A | B | C | D |
| A▶ funds | B▶ costs | C▶ fees | D▶ rates | | | | |

Question 51– 60

- ❖ In this section you must choose the word or phrase which best completes each sentence.
- ❖ For questions **51** to **60**, mark one letter **A**, **B**, **C** or **D** on your Answer Sheet.

| | | | | | | | |
|--|-----------------------|---------------------|--------------------|----------|----------|----------|----------|
| 51. If you're not too tired we could have a.....of tennis after lunch. | | | | A | B | C | D |
| A▶ match | B▶ play | C▶ game | D▶ party | | | | |
| 52. Don't you get tired.....watching TV every nigh? | | | | A | B | C | D |
| A▶ with | B▶ by | C▶ of | D▶ at | | | | |
| 53. Go on, finish the dessert. It needs.....up because it won't stay fresh until. | | | | A | B | C | D |
| A▶ eat | B▶ eating | C▶ to eat | D▶ eaten | | | | |
| 54. We're not used to.....invited to very formal occasions. | | | | A | B | C | D |
| A▶ be | B▶ have | C▶ being | D▶ having | | | | |
| 55. I'd rather we.....meet this evening, because I'm very tired. | | | | A | B | C | D |
| A▶ wouldn't | B▶ shouldn't | C▶ hadn't | D▶ didn't | | | | |
| 56. She obviously didn't want to discuss the matter so I didn't.....the point. | | | | A | B | C | D |
| A▶ maintain | B▶ chase | C▶ follow | D▶ pursue | | | | |
| 57. Anyone.....after the start of the play is not allowed in until the interval. | | | | A | B | C | D |
| A▶ arrives | B▶ has arrived | C▶ arriving | D▶ arrived | | | | |
| 58. This new magazine iswith interesting stories and useful information. | | | | A | B | C | D |
| A▶ full | B▶ packed | C▶ thick | D▶ compiled | | | | |
| 59. The restaurant was far too noisy to be.....to relaxed conversation. | | | | A | B | C | D |
| A▶ conducive | B▶ suitable | C▶ practical | D▶ fruitful | | | | |
| 60. In this branch of medicine, it is vital toopen to new ideas. | | | | A | B | C | D |
| A▶ stand | B▶ continue | C▶ hold | D▶ remain | | | | |

Appendix 11: Elicited Imitation Task

Student Name: _____

| | The sentence | score |
|-------|---|--------------|
| 1 | The red book is on the table. | |
| 2 | I doubt that he knows how to drive that well. | |
| 3 | After dinner, I had a long, peaceful nap. | |
| 4 | The houses are very nice but too expensive. | |
| 5 | The little boy whose kitten died yesterday is sad. | |
| 6 | You really enjoy listening to country music, don't you? | |
| 7 | Cross the street at the light and then just continue straight ahead. | |
| 8 | A good friend of mine always takes care of my neighbour's three children. | |
| 9 | The terrible thief whom the police caught was very tall and thin. | |
| 10 | There are a lot of people who don't eat anything at all in the morning. | |
| Total | | |

Scoring Rubric

| Marking Criteria | score |
|---|--------------|
| Perfect repetition | 4 |
| Accurate content repetition with some (ungrammatical or grammatical) changes. | 3 |
| Changes in content or in form that affect meaning | 2 |
| Repetition of half of the stimulus or less | 1 |
| Silence, only one word repeated, or unintelligible repetition | 0 |

Appendix 12: Working memory tests (Arabic & English)

إختبار الذاكرة العاملة (العد العكسي)

هذا الإختبار مصمم لقياس مدى الذاكرة العاملة من ناحية التخزين والمعالجة لدى المشاركين في البحث. يطلب من المشاركين الإستماع إلى مجموعات من الأرقام وإعادتها بشكل عكسي. عدد الأرقام لكل مجموعة سوف تزداد في كل مرة وسوف تقدم من خلال ملف صوتي وبفاصل زمني ثانية واحدة بين كل رقم. مدى الذاكرة العاملة يتحدد بناءً على آخر مجموعة من الأرقام تم إعادتها بنجاح مرتين.

تعليمات الإختبار

سوف تستمع لمجموعات مختلفة من الأرقام. سوف أذكر الأرقام وأنت تكررهما بشكل عكسي. عدد الأرقام في كل مجموعة سوف يزداد بشكل مطرد. سوف نبدأ بمجموعة من ثلاث أرقام. عندما تحقق محاولتين ناجحتين ننتقل للمجموعة التالية (4 أرقام) وهكذا بشكل متزايد حتى 9 أرقام كحد أعلى. ينتهي الإختبار عندما لا تنجح مرتين في تكرار أي من المجموعات.

مثال:

عندما أقول: 5 4 3

أنت تقول: 3 4 5

أخبرني عندما تكون مستعداً.

| المدى | المحاولة الأولى | √ / x | المحاولة الثانية | √ / x | المحاولة الثالثة | √ / x |
|---------------------|-----------------|-------|------------------|-------|------------------|-------|
| ثلاثة | 524 | | 936 | | 715 | |
| أربعة | 7913 | | 5146 | | 9762 | |
| خمسة | 41527 | | 64951 | | 41539 | |
| ستة | 639514 | | 197249 | | 269721 | |
| سبعة | 3915372 | | 9172631 | | 4962413 | |
| ثمانية | 72529416 | | 53719231 | | 62916473 | |
| تسعة | 173956431 | | 971492564 | | 316497625 | |
| إسم الطالب | | | | | | |
| مدى الذاكرة العاملة | | | | | | |

Backward Digits Span Test (English Language)

This auditory task is designed to test learners' complex working memory capacity (storage and processing). Participants are required to listen to sets of increased digits and repeat them backwards. Numbers are recorded at one digit per second. Each learners' working memory span is determined based on the last digits set he/she has repeated successfully twice.

Instructions:

You are going to listen to different sets of numbers. I will say the numbers and you have to repeat each set **backwards**. Digits will be in increased sets sizes. We will start with sets of three digits. When you have two successful attempts, you move to the next set (4 digits), and so on. The test finishes when you fail twice to repeat any of the sets.

For example:

When I say: "4 5 6"

You say: "6 5 4"

Let me know when you are ready.

| Span | First trial | √ / × | Second trial | √ / × | Third trial | √ / × |
|------------------------------------|-------------|----------|--------------|----------|-------------|-------|
| Three | 582 | | 395 | | 627 | |
| Four | 3915 | | 4826 | | 1973 | |
| Five | 68471 | | 73169 | | 25184 | |
| Six | 592834 | | 469172 | | 358261 | |
| Seven | 7452846 | | 8361957 | | 5913728 | |
| Eight | 92518753 | | 16829374 | | 81492573 | |
| Nine | 483261759 | | 692748315 | | 751936845 | |
| Student name | | | | | | |
| Backward Digits Span result | | | | | | |

Appendix 13: Samples of the coded data (Study One)

| Participant: 4 | | Task: Flying Car (+IR) | | Time: 122sec | | Words: 187 / 152 | |
|--------------------|----|---------------------------------|----|-------------------------|----|------------------|--|
| OPT: A2 | | LP: 52 / 100 | | L1 WM: 4 | | L2 WM: 4 | |
| No of AS-units | 27 | No of Clauses :: | 34 | Abandoned utter X | 2 | | |
| No of errors # | 18 | Error-free clauses errfr | 16 | Filled pauses eh | 71 | | |
| Repetitions * | 18 | Reformulations ~ | 2 | Replacements ∞ | 2 | | |
| False starts “” | 2 | Hesitations Λ | 1 | Dysfluencies | 25 | | |
| Level 1 error (.8) | 10 | Level 2 error (.5) | 7 | Level 3 error (.1) | 1 | | |

1Pat and Mat are driving their car **errfr** | 2I think **errfr** :: 3they will go to ∞ eh in a trip # | 4eh eh it's “” eh eh they stop (1.08) **errfr** | 5eh they look to the roof # | 6I think **errfr** :: 7they will eh (0.44) cut the roof (0.68) **errfr** | 8or eh Mat * eh Mat tells Pat **errfr** :: 9to look it eh to the sky in ∞ to the plane # (1.88) | 10eh eh Pat are cut the eh roof # (1.92) | 11I think **errfr** :: 12eh he will eh make eh wings **errfr** :: 13to eh fly the car ~ to make the car fly **errfr** | 14eh eh (0.57) eh they * they saw **errfr** :: 15that the car eh is eh try to * to fly # (0.66) | 16they stop another eh # X | 17it was ~ eh it is heavy **errfr** | 18they stop again (1.15) **errfr** | 19eh they eh take away their eh bags # | 20eh and * eh eh (1.25) and eh eh eh eh (0.90) drive again **errfr** | 21eh it * eh it still eh heavy the car # | 22eh eh I think **errfr** :: 23they will stop again **errfr** | 24then they will eh take out the eh motor # | 25they * they fix the motor # | 26eh they * they * eh they eh give out the * eh the motor from the car # | 27eh and sta Λ and eh eh switch on the eh * switch on the car # | 28it doesn't * eh eh it doesn't eh eh eh eh # X | 29it was a truck eh # | 30it come again # | 31eh he * he eh put * eh eh he put eh eh the car in the eh truck # | 32eh the * eh the “” he * he catch * he catch the * eh he catch the * eh eh the eh car # | 33the car is flying now **errfr** | 34eh eh he put a glass # |

| Participant: 4 | Flying Car (+IR) | LP: 52% (B1) | WM: 4 |
|-----------------------------------|--|---------------------|--------------------|
| Dimensions | Measures | | Results |
| Syntactic Complexity | Mean length of AS-units | | 6.92 / 5.62 |
| | Mean length of clauses | | 5.50 / 4.47 |
| | Ratio of subordination | | 1.25 |
| Lexical Complexity | D | | 11.69 |
| | Plex Lambda | | 3.68 / 1.05 |
| Accuracy | Percentage of error free clauses | | 47.05 |
| | Weighted Clause Ratio (WCR) | | .81 |
| Fluency | Speech rate (words per minute) | | 91.9 / 74.7 |
| | Number of mid-clause silent pauses (≥ 0.40 sec) | | 4 |
| | Number of end-clause silent pauses (≥ 0.40 sec) | | 6 |
| | Mean length of mid-clause silent pauses (≥ 0.40 sec) | | .57 |
| | Mean length of end-clause silent pauses (≥ 0.40 sec) | | 1.22 |
| | Number of filled pauses per minute | | 34.91 |
| | Number of repetitions per minute | | 5.85 |
| | Number of reformulations per minute | | .98 |
| | Number of replacements per minute | | .98 |
| | Number of false starts per minute | | .98 |
| | Number of hesitations per minute | | .49 |
| Number of dysfluencies per minute | | 12.29 | |

| | | | | | |
|--------------------|---------------------|---------------------------------|------------------|--------------------------|-----------|
| Participant: 4 | Task: Lunch (-IR) | Time: 125 sec | Words: 172 / 135 | | |
| OPT: A2 | LP: 52 / 100 | L1 WM: 4 | L2 WM: 4 | | |
| No of AS-units | 27 | No of Clauses :: | 31 | Abandoned utter ⌘ | 1 |
| No of errors # | 20 | Error-free clauses errfr | 11 | Filled pauses eh | 44 |
| Repetitions * | 9 | Reformulations ~ | 6 | Replacements ∞ | 4 |
| False starts “” | 0 | Hesitations Λ | 4 | Dysfluencies | 23 |
| Level 1 error (.8) | 9 | Level 2 error (.5) | 10 | Level 3 error (.1) | 1 |

1Pat is listening to music **errfr** | **2**and Mat eh is reading a book **errfr** | **3**eh Mat is standing up **errfr** | **4**and look * look to the watch # (0.62) | **5**eh (1.55) eh eh Mat is co **Λ** is cooking the food **errfr** (1.45) | **6**eh he saw chicken apple # (1.09) eh | **7**he make a fire # (0.55) | **8**he want # :: **9**to make a fire **errfr** (2.71) | **10**he eh put a water # (2.43) | **11**eh he give eh food ~ (0.62) a type of food # (0.80) | **12**then he eh eh make a water ∞ a fire # (1.69) | **13**Mat give ~ eh Mat (0.89) eh help eh Pat # (0.69) | **14**eh Mat eh make a fire # (2.99) | **15**eh Pa **Λ** Pa **Λ** ah there * eh there is eh eh a leaking water # | **16**it's rain # | **17**the sky is * is rain # (0.86) | **18**eh eh Mat and ~ Mat give ~ Pat give umbrella # (1.00) | **19**eh Mat ∞ eh he * he told Pat # **⌘** | **20**Mat didn't give umbrella # :: **21**inside the ~ eh to go inside the * eh the house **errfr** | **22**eh they put ~ they go inside the house **errfr** | **23**and eh eh they put * they * they put the food in the eh oven **errfr** | **24**eh he put a wood in the oven # :: **25**to eh make a fire **errfr** | **26**eh eh there's a leaking in the roof # | **27**eh eh Pat ∞ he saw the roof **errfr** | **28**it's eh lea **Λ** it's leaking **errfr** | **29**eh he give a eh eh eh chair # | **30**eh he put a * eh a paper ∞ a tissue * a tissue # :: **31**to don't the leaking # |

| Participant: 4 | Lunch (-IR) | LP: 52% (B1) | WM: 4 |
|-----------------------------|--|---------------------|--------------------|
| Dimensions | Measures | | Results |
| Syntactic Complexity | Mean length of AS-units | | 6.37 / 5.0 |
| | Mean length of clauses | | 5.54 / 4.35 |
| | Ratio of subordination | | 1.14 |
| Lexical Complexity | D | | 14.09 |
| | Plex Lambda | | 4 / 1.82 |
| Accuracy | Percentage of error free clauses | | 35.48 |
| | Weighted Clause Ratio (WCR) | | .75 |
| Fluency | Speech rate (words per minute) | | 82.5 / 64.8 |
| | Number of mid-clause silent pauses (≥ 0.40 sec) | | 4 |
| | Number of end-clause silent pauses (≥ 0.40 sec) | | 11 |
| | Mean length of mid-clause silent pauses (≥ 0.40 sec) | | 1.03 |
| | Mean length of end-clause silent pauses (≥ 0.40 sec) | | 1.43 |
| | Number of filled pauses per minute | | 21.12 |
| | Number of repetitions per minute | | 4.32 |
| | Number of reformulations per minute | | 2.88 |
| | Number of replacements per minute | | 1.92 |
| | Number of false starts per minute | | 0 |
| | Number of hesitations per minute | | 1.92 |
| | Number of dysfluencies per minute | | 11.04 |

| | | | | | |
|--------------------|------------------------|---------------------------------|------------------|--------------------------|-----------|
| Participant: 24 | Task: Flying Car (+IR) | Time: 127 sec | Words: 267 / 220 | | |
| OPT: B2 | LP: 76 / 100 | L1 WM: 5 | L2 WM: 5 | | |
| No of AS-units | 26 | No of Clauses :: | 48 | Abandoned utter X | 0 |
| No of errors # | 9 | Error-free clauses errfr | 39 | Filled pauses eh | 35 |
| Repetitions * | 11 | Reformulations ~ | 4 | Replacements ∞ | 1 |
| False starts “” | 0 | Hesitations Δ | 0 | Dysfluencies | 16 |
| Level 1 error (.8) | 9 | Level 2 error (.5) | 0 | Level 3 error (.1) | 0 |

1Pat and Mat are in a car **errfr** (0.88) | **2**eh it seems like **errfr** :: **3**eh they are going on a picnic **errfr** :: **4**because they have luggage **errfr** (0.90) I think (0.95) | **5**now they stop for a random reason # (1.22) | **6**I think **errfr** :: **7**they are trying to realise **errfr** :: **8**what's the problem # (0.97) | **9**eh then eh Mat eh (0.66) eh stare # :: **10**eh shows Pat eh a ∞ an airplane **errfr** (2.06) | **11**eh then they are cutting eh the roof **errfr** (1.06) :: **12**so they can fly **errfr** (2.65) | **13**eh they are trying to fly **errfr** (0.45) | **14**but there is some kind of problem # (0.94) :: **15**because the car is heavy **errfr** (1.01) | **16**eh (1.07) till now they are not realising the problem **errfr** | **17**then (0.75) they stare at each other **errfr** | **18**then (0.57) Pat * eh (0.79) then Pat looks at the luggage **errfr** (0.85) | **19**they decide **errfr** :: **20**to throw the luggage **errfr** :: **21**to make the eh car lighter **errfr** (1.57) | **22**and that's (1.61) ~ it's may be # (0.59) :: **23**it * it can be worth it **errfr** (0.59) | **24**if they want something **errfr** :: **25**they can do anything **errfr** :: **26**to do it **errfr** | **27**I think **errfr** :: **28**it's worth it **errfr** | **29**then eh it's not working too **errfr** | **30**they are looking at the eh engine **errfr** | **31**they open eh the car # :: **32**to see the engine **errfr** | **33**and then they start **errfr** :: **34**fixing it **errfr** | **35**they * eh they take the motor out # | **36**and eh try ~ and trying ~ they are trying eh to * eh eh to eh they * they are trying to * eh they * they are trying to **errfr** :: **37**turn on the car # | **38**but it's not working **errfr** :: **39**because it ~ because they throw * eh they throw the engine # | **40**then eh they get the tow car **errfr** :: **41**eh to * eh to pull eh to pull * their car **errfr** | **42**eh and then they are kind of flying **errfr** | **43**eh eh eh they are flying **errfr** :: **44**because of their decisions **errfr** | **45**and that's * and that's why **errfr** :: **46**eh you have to do some eh sacrifices **errfr** :: **47**to reach something **errfr** :: **48**you want **errfr** |

| Participant: 24 | Flying Car (+IR) | LP: 76% (B2) | WM: 5 |
|-----------------------------------|--|---------------------|---------------------|
| Dimensions | Measures | | Results |
| Syntactic Complexity | Mean length of AS-units | | 10.26 / 8.46 |
| | Mean length of clauses | | 5.56 / 4.58 |
| | Ratio of subordination | | 1.84 |
| Lexical Complexity | D | | 25.10 |
| | Plex Lambda | | 2.22 / .80 |
| Accuracy | Percentage of error free clauses | | 81.25 |
| | Weighted Clause Ratio (WCR) | | .96 |
| Fluency | Speech rate (words per minute) | | 126 / 103.9 |
| | Number of mid-clause silent pauses (≥ 0.40 sec) | | 7 |
| | Number of end-clause silent pauses (≥ 0.40 sec) | | 14 |
| | Mean length of mid-clause silent pauses (≥ 0.40 sec) | | .90 |
| | Mean length of end-clause silent pauses (≥ 0.40 sec) | | 1.12 |
| | Number of filled pauses per minute | | 16.53 |
| | Number of repetitions per minute | | 5.19 |
| | Number of reformulations per minute | | 1.88 |
| | Number of replacements per minute | | .47 |
| | Number of false starts per minute | | 0 |
| | Number of hesitations per minute | | 0 |
| Number of dysfluencies per minute | | 7.55 | |

| | | | | | |
|--------------------|---------------------|---------------------------------|-------------------|-------------------------|-----------|
| Participant: 24 | Task: Lunch (-IR) | Time: 123 sec | No. of words: 242 | | |
| OPT: B2 | LP: 76 / 100 | L1 WM: 5 | L2 WM: 5 | | |
| No of AS-units | 33 | No of Clauses :: | 43 | Abandoned utter X | 1 |
| No of errors # | 19 | Error-free clauses errfr | 24 | Filled pauses eh | 27 |
| Repetitions * | 8 | Reformulations ~ | 3 | Replacements ∞ | 2 |
| False starts “” | 1 | Hesitations Λ | 2 | Dysfluencies | 16 |
| Level 1 error (.8) | 18 | Level 2 error (.5) | 1 | Level 3 error (.1) | 0 |

1Pat is listening to some music **errfr** | 2and he is sewing **errfr** (0.70) | 3then Mat is eh checking the time **errfr** (0.46) | 4then he shows eh Pat # X (0.83) | 5eh he's checking the table **errfr** (1.26) | 6eh breakfast ~ he's (0.90) wants # :: 7to see the breakfast # (1.80) | 8eh (0.78) he's looking at the breakfast # | 9and wants to do something **errfr** (0.48) | 10ah he's making breakfast **errfr** (0.64) | 11eh he's pouring water into a pot and lemons eggs # (0.62) | 12and he is making a fire **errfr** (0.70) | 13eh Pat comes **errfr** :: 14to helps him * to help him **errfr** (0.72) | 15eh (0.91) Mat has a lighter # (0.99) | 16he fires it # (0.42) | 17it eh turns off # (1.01) | 18eh (0.94) Pat “” (0.99) ah there are drop ∞ drops of water coming **errfr** (0.52) | 19eh and Pat * (0.91) Pat realises **errfr** :: 20that it's raining **errfr** (0.58) | 21eh they both * (0.48) they bo Λ they both pick an umbrella **errfr** (0.82) :: 22to put it off the breakfast # | 23and sit under it # | 24so they can't be wet # | 25Pat gets into the house **errfr** | 26and moves the eh pots and breakfast to the house **errfr** | 27eh Ma Λ Mat follows him to the house **errfr** | 28eh he puts the eggs ~ eh Pat * Pat puts the eggs eh in the oven **errfr** | 29and * eh and Mat helps him **errfr** :: 30with putting eh some wood in the oven # :: 31to eh fire it # | 32eh he gets the lighter # :: 33to light ∞ eh to make fire **errfr** | 34then a drop of water eh from the ~ eh from a hole in the roof comes **errfr** :: 35to turn off the lighter # | 36then they see the drop **errfr** | 37and eh know where it is **errfr** | 38so Pat wants **errfr** :: 39to * eh to close it # | 40and he picks up a tissue **errfr** :: 41to eh close the hole by the help of Mat # | 42then * then Mat * then Mat lights the lighter # | 43and make breakfast # |

| Participant: 24 | Lunch (-IR) | LP: 76% (B2) | WM: 5 |
|-----------------------------------|--|---------------------|--------------------|
| Dimensions | Measures | | Results |
| Syntactic Complexity | Mean length of AS-units | | 7.33 / 6.63 |
| | Mean length of clauses | | 5.62 / 5.09 |
| | Ratio of subordination | | 1.30 |
| Lexical Complexity | D | | 30.9 |
| | Plex Lambda | | 2.88 / 1.58 |
| Accuracy | Percentage of error free clauses | | 55.81 |
| | Weighted Clause Ratio (WCR) | | .90 |
| Fluency | Speech rate (words per minute) | | 118 / 106.8 |
| | Number of mid-clause silent pauses (≥ 0.40 sec) | | 7 |
| | Number of end-clause silent pauses (≥ 0.40 sec) | | 16 |
| | Mean length of mid-clause silent pauses (≥ 0.40 sec) | | 0.84 |
| | Mean length of end-clause silent pauses (≥ 0.40 sec) | | 0.78 |
| | Number of filled pauses per minute | | 13.17 |
| | Number of repetitions per minute | | 3.90 |
| | Number of reformulations per minute | | 1.46 |
| | Number of replacements per minute | | 0.97 |
| | Number of false starts per minute | | 0.48 |
| | Number of hesitations per minute | | 0.97 |
| Number of dysfluencies per minute | | 7.80 | |