How stable is second language aptitude? Effects of second language learning and language analysis training on second language aptitude test scores

James Chalmers

BA (Hons.) in Languages & Applied Linguistics

School of Humanities, Languages and Social Science

Arts, Education & Law Group

Griffith University

Submitted in fulfilment of the requirements of the degree of Doctor of Philosophy

December 2017
While second language (L2) aptitude has traditionally been considered a stable individual factor in SLA, more recent research and theory has questioned this assumption (Singleton, 2017). If L2 aptitude is stable, then the implication is that this set of cognitive abilities are genetic in origin and/or limited (Skehan, 1998). On the other hand, if L2 aptitude is not stable and is sensitive to experience, then it suggests that whatever constitutes L2 aptitude can be taught to make initial L2 learning easier and faster. While more recent studies have found that the L2 learning experience itself seems to have a training effect on L2 aptitude test scores (Ganschow, 1993; Ganschow & Sparks, 1995; Sáfár & Kormos, 2008; Sparks, Ganschow, Pohlman, Skinne, & Artzer, 1992; Sparks, Ganschow, Artzer, & Patton, 1997), both the size of such training effects and the effects of direct training of L2 aptitude abilities (Politzer & Weiss, 1969) remain unclear.

The current study contributes to this debate by investigating the stability of L2 aptitude test scores. Conducted over an 8-week period of L2 instruction, the study investigated whether 6 weeks of aptitude training would enhance language analytic abilities, an area less researched in previous studies. Participants were 85 university students taking an introductory Spanish course, which included 6 weeks of instruction and practice on language analysis for learning L2 Spanish. The study operationalised L2 aptitude in terms of the abilities measured by the LLAMA tests: associative memory (LLAMA B), sound discrimination (LLAMA D), sound-symbol association (LLAMA E), and language analytic abilities (LLAMA F). Pre-testing took place at the beginning of L2 instruction and post-testing was conducted after the completion of the language analysis skills instruction. Stability was investigated through changes in L2 aptitude test scores. Results showed that post-test scores were significantly
higher for all LLAMA tests except for the LLAMA E (sound-symbol association), which showed a ceiling effect in both pre- and post-tests. At the whole-group level, the size of gain scores differed across all tests with the greatest gains on the LLAMA B (associative memory) and the smallest for the LLAMA F (language analytic abilities). However, at the sub-group level, it was mainly participants with lower than average pre-test scores who achieved significant gains, with comparable effect sizes across all tests (except the LLAMA E).

Trainability was investigated by comparing the predictive ability of training variables (e.g. accuracy scores and speed on training tasks) with L2 aptitude pre-test scores. Results for the predictive ability of training language analytic abilities were mixed on post-test scores. A random-forest regression (see Strobl, Malley, & Tutz, 2009) found that accuracy scores on the training were more predictive of post-test scores than LLAMA pre-test scores, but a step-wise linear regression did not confirm this finding. However, training variables were important predictors in both regression models.

Overall, the findings suggest that L2 aptitude scores increase over time with L2 learning experience and instruction, in line with previous longitudinal studies (Sparks, Ganschow, Artzer, & Patton, 1997). This may imply that L2 aptitude is trainable, with some training variables being important predictors of post-test scores. The sensitivity of L2 aptitude measures to L2 learning experience and instruction may have implications for the validity of static, one-off measures of L2 aptitude for predicting L2 achievement.
ACKNOWLEDGEMENTS

A PhD is a long journey and the number of people who have helped me along the way is large, to say the least. I owe them all a debt of gratitude. I would like to first say thanks to my supervisors: Dr Susana Eisenchlas, Dr Andrea Schalley, and Dr Andrew Munro – all from the School of Humanities, Languages, and Social Sciences at Griffith University. For your unwavering support, your guidance, your questions, and your infinite understanding I can never thank you enough. It would be foolish to try and quantify the contribution you have made to my life, but above all I need to thank you for your belief in me. Without this, I would never have achieved half of what I did. For your endless hours and expertise, you deserve a lot more than I can give. Rest assured, I am forever in your debt.

I would also like to thank Dr Jana Hart'anská for her supervision and support early on in the project. For taking on a complete stranger in a strange land, I would like to express my thanks. Thank you for your guidance, your effort, and your help.

My warmest thanks need to go to Dr Bill Venables from the Mathematics, Informatics and Statistics division of the Commonwealth Scientific and Industrial Research Organisation (CSIRO). Your infinite patience, wisdom, and generosity were an unexpected gift during this project. I am truly grateful for you introducing me to the world of statistics and R, despite the fact that I thought it might break my brain on more than one occasion!

Above all others, I need to thank my partner in crime, my beautiful wife, Mishka. For being my rock, for never giving up on me, and for always checking in on how ‘our’ PhD was going, I do not have enough words to express my love for you. And to my amazing twins, Mishko and Sasha, thank you for always being there to take my mind off things when I needed a
break, for filling my life with love, and showing me that the world is bigger than my research project. This thesis is for the three of you, because without you, it would never have been possible.

Lastly, I would like to thank all the people who helped me along the way. The list is too long to write in full, but it is in my head and I will think of you all whenever I reflect back on this time. Cheers!
Statement of originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

James Chalmers

Ethical clearance

This project received ethical clearance from the Griffith University Human Research Ethic Committee under protocol number LAL/05/14/HREC.
# TABLE OF CONTENTS

## CHAPTER 1: INTRODUCTION

1.1 Context of the current study 19  
1.2 Theoretical framework 23  
1.3 Aims of the study 24  
1.4 Overview of current study 26  
1.5 Contribution to the field 26  
1.6 Outline of chapters 28

## CHAPTER 2: WHAT IS SECOND LANGUAGE APTITUDE

2.1 Definitions and theories of L2 aptitude 30  
2.1.1 The MLAT as a proxy for L2 aptitude 31  
2.1.2 Carroll’s four-factor construct of L2 aptitude 35  
2.1.3 Skehan’s (2002) model of L2 aptitude and SLA 40  
2.1.4 Robinson’s Aptitude Complex Hypothesis 43  
2.1.5 Cognitive Ability for Novelty in Acquisition of Language (Foreign) Test (CANAL-FT) Construct 48  
2.1.6 L2 aptitude theory: Conclusions 52  
2.2 Systematic review of L2 aptitude research 53  
2.2.1 Procedure for systematic review 54  
2.2.2 Study selection 56  
2.2.3 Terminology 57  
2.2.4 Results: The general characteristics of L2 aptitude 58  
2.2.4.1 Publication details of L2 aptitude research 58  
2.2.4.2 Geographical spread 61  
2.2.4.3 Research methods 64
CHAPTER 3: INVESTIGATING THE STABILITY OF SECOND LANGUAGE APTITUDE AND THE TRAINABILITY OF LANGUAGE ANALYTIC ABILITIES 91

3.1 Stability and untrainability of second language aptitude 95
   3.1.1 Carroll’s theoretical assumptions 95
   3.1.2 Empirical studies of stability of L2 aptitude 97
      3.1.2.1 Pre- and post-test studies of L2 aptitude 99
      3.1.2.2 Cross-sectional studies of L2 aptitude and L2 learning experience 101
      3.1.2.3 Training studies of L2 aptitude 105
      3.1.2.4 Summary of evidence on stability & untrainability of L2 aptitude 109
   3.1.3 Alternatives to Carroll’s perspective of L2 aptitude 110
      3.1.3.1 Second language aptitude is based on general cognitive abilities 111
3.1.3.2 Second language aptitude as a hierarchical system 112
3.1.3.3 Second language learners as adaptive 115
3.1.3.4 Strategies and knowledge of second language learning can be taught 116
3.1.3.5 Summary of alternatives to Carroll’s perspective of L2 aptitude 118
3.1.4 Summary of stability and trainability of L2 aptitude 119

3.2 Language analytic abilities: a training perspective 120
   3.2.1 Terms, definitions, and issues with language analytic abilities 121
   3.2.2 What we know about language analytic abilities: tests and empirical studies 122
       3.2.2.1 Measures of language analytic abilities 122
       3.2.2.2 Studies of language analytic abilities 128
       3.2.2.3 Summary 131
   3.2.3 A novel conceptualisation of language analytic abilities 132
   3.2.4 List of potential language analytic abilities 134
   3.2.5 Summary of language analytic abilities and the implications for training 140

3.3 Research questions 141

CHAPTER 4: METHODOLOGY 140

4.1 Research Design 143
4.2 Participants 144
   4.2.1 Ethical considerations 149
4.3 Procedure 150
4.4 Materials 153
   4.4.1 L2 aptitude test instrument 153
   4.4.2 Language analytic abilities training 159
   4.4.3 The training tasks 161
       4.4.3.1 Pattern Matching 162
       4.4.3.2 Problem Solving 163
CHAPTER 5: ANALYSIS AND RESULTS

5.1 Research question 1.a: Are post-test scores for L2 aptitude, operationalised as the LLAMA tests, significantly different to pre-test scores? 194

5.1.1 Participant flow and recruitment 194
5.1.2 Assumptions of paired-samples t-tests 195
5.1.3 Background variables of participants 201
5.1.4 Descriptive statistics  
5.1.5 Paired-samples t-tests (no groups)  
5.1.6 Paired-sample t-tests for low and high groups  
5.1.7 Pre- and post-test correlations  
5.1.8 Summary for research question 1  

5.2 Research question 2: Do training factors explain more variance in post-test scores for language analytic abilities than L2 aptitude factors?  
5.2.1 Data preparation  
  5.2.1.1 Trimming times for training tasks and LLAMA tests  
  5.2.1.2 Initial aptitude level  
  5.2.1.3 Participants’ approach to training  
  5.2.1.4 Training accuracy performance  
  5.2.1.5 Training time durations performance  
5.2.2 Random forest regression model  
  5.2.2.1 Stepwise linear regression follow-up  
5.2.3 Summary for research question 2  

5.3 Summary of results for the research questions  

CHAPTER 6: DISCUSSION  

6.1 Research question 1.a: Do post-test scores for L2 aptitude differ significantly from pre-test scores?  
  6.1.1 Pre- to post-test gains  
  6.1.2 Size of gains  
  6.1.3 LLAMA F differences for L1-L2 level  
  6.1.4 Test-retest reliability  
  6.1.5 Summary  

6.2 Research questions 1.b: Do training factors explain more variance in post-test scores for language analytic abilities than L2 aptitude factors?
6.2.1 Predictive power of random forest model 252
6.2.2 Relative importance of predictors 254
6.2.3 Linear model in comparison 255
6.2.4 Summary 256

6.3 Implications of the study 257
6.3.1 Test training effect 259
6.3.2 L2 experience as training effect 260

6.4 Challenges and limitations of study 263

CHAPTER 7: CONCLUSION 263

7.1 Main findings of the study 266
7.2 Implications of findings 269
7.3 Limitations of the research 270
7.4 Recommendations for future work 271

APPENDIX A: PROCEDURE FOR THE SYSTEMATIC REVIEW OF L2 APTITUDE RESEARCH 272

A.1 Objectives 274
A.2 Process 274
  A.2.1 Data items 276
  A.2.2 Study selection 278
  A.2.3 Terminology 279
  A.2.4 Data items and their definitions 280

APPENDIX B: REFERENCE LIST FOR PAPERS IN SYSTEMATIC REVIEW OF L2 APTITUDE 295
APPENDIX C: SUPPLEMENTARY RESULTS

C.1 Introduction

C.2 Research Question 1: Are post-test scores for L2 aptitude, operationalised as the LLAMA tests, significantly different to pre-test scores?
   C.2.1 Assumptions of paired-samples t-tests
   C.2.2 Background variables of participants
   C.2.3 Descriptive statistics
   C.2.4 Paired-samples t-tests
      C.2.4.1 Paired-sample t-tests for low and high groups
      C.2.4.2 Pre- and post-test correlations

C.3 Research Question 2: Do training factors explain more variance in post-test scores for language analytic abilities than L2 aptitude factors?
   C.3.1 Data preparation
      C.3.1.1 Trimming times for training tasks and LLAMA tests
      C.3.1.2 Initial aptitude level
      C.3.1.3 Approach to training course
      C.3.1.4 Training task accuracy
      C.3.1.5 Training task speed
   C.3.2 Random forest regression model
      C.3.2.1 Stepwise linear regression follow up
LIST OF TABLES

2.1 Carroll’s (1962:129-130) four factors of L2 aptitude 35
2.2 The sub-tests of the MLAT and the factors they purportedly measure (Carroll, 1962; Sasaki, 2012) 36
2.3 SLA processing stages and potential aptitude components (Skehan, 2002:90) 40
2.4 Key to Figure 2.2 (Robinson 2005:53) 45
2.5 CANAL-FT acquisition processes 49
2.6 CANAL-FT test tasks and their cognitive processes and levels of processing 50
2.7 The prevalence of key research themes in L2 aptitude research 69
3.1 Descriptions of the MLAT IV, PLAB 4, and LLAMA F measures of language analytic abilities 123
4.1 Descriptive statistics for age of participants 145
4.2 Participants’ first language backgrounds 145
4.3 L2 experience levels of participants 146
4.4 Summary of training task types and their relationship to language analytic abilities 173
4.5 Grammatical gender of nouns in Spanish (Whitley, 1986) 185
4.6 Syllabus overview of language analytic abilities training course 191
5.1 Skewness, kurtosis, and Shapiro-Wilks for LLAMA tests 198
5.2 LLAMA F pre-test scores for English vs non-English L1 participants for differing levels of L2 experience 204
5.3 Descriptive statistics for all LLAMA pre- and post-tests 206
5.4 Results from paired-sample t-tests, 95% confidence intervals, and effect sizes (Cohen’s d) for each LLAMA test (B, D, E, F, and Total) 207
5.5 Results from paired-sample t-tests for Low and High groups for all LLAMA tests 213
5.6 Pearson correlations between pre- and post-tests for all LLAMA tests 217
5.7 The 23 variables included in the random forest regression model 226
5.8 Results of random forest regression model predicting LLAMA F post-test scores 231
5.9 The 18 variables included in the stepwise linear regression model 233
5.10 Pearson correlations for predictor variables (listed) with response variable of LLAMA F post-test scores 237

5.11 Final model from stepwise regression predicting LLAMA F post-test scores 238

A.1. Keywords used in database search 275

A.2. Themes and questions for L2 aptitude research 276

A.3. Data items and their definitions 280

C.1 Skewness, kurtosis, and Shapiro-Wilks for LLAMA tests 313

C.2 LLAMA F pre-test scores for English vs non-English L1 participants for differing levels of L2 experience 322

C.3 Descriptive statistics for all LLAMA pre- and post-tests 323

C.4 Results from paired-sample t-tests, 95% confidence intervals, and effect sizes (Cohen’s d) for each LLAMA test (B, D, E, F, and Total) 324

C.6 Pearson correlations between pre- and post-tests for all LLAMA tests 331

C.7 Results of random forest regression model predicting LLAMA F post-test scores 339

C.8 Pearson correlations for predictor variables (listed) with response variable of LLAMA F post-test scores 341

C.9 Final model from stepwise regression predicting LLAMA F post-test scores 353

LIST OF FIGURES

2.1 Example question from the MLAT IV Words in Sentences sub-test 38

2.2 Aptitudes, development, and learning contexts: Changes in the relative contribution of aptitude factors to different aspects of L2 learning. (Inner two circles: initial input-based learning; third circle: output practice and complex task performance; and outer circle: transfer of task performance to real world interactive settings.) (Robinson, 2005: 52) 45

2.3 Flowchart of screening process for systematic review. 57

2.4 Journals of L2 aptitude research with more than one publication (n > 1) 59

2.5 Number of publications on L2 aptitude categorised by publication type 60

2.6 First authors of L2 aptitude research with more than one paper (n > 1) 61

2.7 Density map of countries where L2 aptitude research has been conducted 62

2.8 Density map of cities where L2 research has been conducted 63
2.9 Density map of countries where data were collected for L2 aptitude studies
64
2.10 Dependent variables in L2 aptitude studies
68
2.11 Independent variables in L2 aptitude studies
69
3.1 Example question from the MLAT IV Words in Sentences sub-test (LLTF, 2013)
124
3.2 Example question from the PLAB 4 Language analysis sub-test (LLTF, 2013)
126
3.3 Example question from the LLAMA F grammatical inference test (Meara, 2005)
126
3.4 Speculative model of language analytic abilities in acquiring L2 knowledge
138
4.1 LLAMA B test console
156
4.2 LLAMA D test console
157
4.3 LLAMA E test console
158
4.4 LLAMA F test console
159
4.5 Speculative model of language analytic abilities in acquiring L2 knowledge
162
4.6 Pattern matching training task example
163
4.7 Problem solving training task example
164
4.8 Grammaticality judgement training task example
165
4.9 Multiple choice training task example
166
4.10 Processing input explanation and summaries example from training
168
4.11 Input processing practice training task example
169
4.12 Google Translate training task example
170
4.13 Example of one type of feature description training task
171
4.14 Q&A training task example
172
4.15 Example of text enhancement in training task (from figure 4.6)
178
5.1 Histograms of LLAMA tests B, D, E, and F
200
5.2 Histogram of LLAMA Total
201
5.3 QQ plots for LLAMA tests B, D, E, F
201
5.4 QQ plot for LLAMA Total
202
5.5 Boxplots for LLAMA B, D, E, F
202
5.6 Boxplot for LLAMA Total
203
5.7 Boxplots for pre- and post-tests for LLAMA B, D, E, F

5.8 Boxplots for pre- and post-tests for LLAMA Total

5.9 Scatter plots for LLAMA B, D, E, F with the regression line (red) and its 95% confidence interval (blue shading)

5.10 Scatter plot for LLAMA Total with the regression line (red) and its 95% confidence interval (blue shading)

5.11 Scatter plots for Low (triangle) and High (circle) groups for LLAMA B, D, E, F with the regression line and its 95% confidence interval (shading)

5.12 Scatter plot for Low (triangle) and High (circle) groups for LLAMA Total with the regression line and its 95% confidence interval (shading)

5.13 Plot of the relative importance of the variables in the random forest regression model (see table 5.7 for variable terms)

A.1 Papers on L2 aptitude research published 1962-2016

A.2 Flowchart of screening process for systematic review.

C.1 Histograms of LLAMA tests B, D, E, and F

C.2 Histogram of LLAMA Total

C.3 QQ plots for LLAMA tests B, D, E, F

C.4 QQ plot for LLAMA Total

C.5 Boxplots for LLAMA B, D, E, F

C.6 Boxplot for LLAMA Total

C.7 Boxplots for pre- and post-tests for LLAMA B, D, E, F

C.8 Boxplots for pre- and post-tests for LLAMA Total

C.9 Scatter plots for LLAMA B, D, E, F with the regression line (red) and its 95% confidence interval (blue shading)

C.10 Scatter plot for LLAMA Total with the regression line (red) and its 95% confidence interval (blue shading)

C.11 Scatter plots for Low (triangle) and High (circle) groups for LLAMA B, D, E, F with the regression line and its 95% confidence interval (shading)

C.12 Scatter plot for Low (triangle) and High (circle) groups for LLAMA Total with the regression line and its 95% confidence interval (shading)
CHAPTER 1: Introduction

This thesis was motivated by my experiences as a language teacher and researcher. As a language teacher, I had observed that some students showed a clear talent for language learning, while others struggled. Yet I had also observed that many students, regardless of any talent, quickly improved their language learning by developing effective second language (L2) learning skills and strategies. As a researcher, I wanted to discover what it was that the most talented language learners knew and did, then explore whether what they knew and did could be taught. Put otherwise, I wanted to train individuals in L2 aptitude to improve their language learning skills.

1.1 Context of the current study

In the field of second language acquisition (SLA), the talent for language learning is known as L2 aptitude, which is often described as the comparative rate and ease with which an individual learns a second language. Although intuitively appealing, this definition lacks any real explanatory power. In reality, a range of individual differences, such as motivation and working memory, contribute to the rate and ease with which one learns a second language. L2 aptitude is, more accurately, the individual difference measured by L2 aptitude tests, most of which consist of problem-solving tasks that purportedly reflect L2 learning. Consequently, before I could teach and potentially train L2 aptitude, I needed to first answer the more fundamental questions of: What exactly is L2 aptitude? Can it be taught? And if it can be taught, how can it be trained?

Surprisingly, even after 60 years of research, it remains unclear exactly what L2 aptitude tests measure, although certain assumptions have, of course, been made. Least controversially, L2
aptitude is widely considered to be an “umbrella” term for a composite of abilities central to L2 learning. In general, three major components of L2 aptitude are purportedly measured by L2 aptitude tests: phonological abilities, memory, and language analytic abilities. Importantly, L2 aptitude tests are assumed to measure the performance of the basic cognitive abilities of these three components, which are involved in L2 learning (Dörnyei, 2005). However, the theory and research on L2 aptitude offer little insight into which specific abilities constitute these components, particularly for language analytic abilities (LAAs henceforth). The considerable research into the Linguistic Coding Deficit Hypothesis in the area of L1 phonological/orthographic skills offers some insights into the nature of L2 aptitude phonological abilities (see Sparks, 1995). Similarly, the research into working memory and its constituent parts is well documented in the fields of L2 aptitude, cognitive psychology, and also SLA. In contrast, LAAs have been studied only indirectly, mainly in relation to the acquisition of grammar in SLA, with little research exploring the nature of this component of L2 aptitude. Considering the importance of LAAs in predicting L2 achievement (Li, 2015), this area of L2 aptitude merits more research.

L2 aptitude is also traditionally assumed to be specific to the phenomenon of language and independent from general intelligence. This assumption is based on stronger correlations between tests of L2 aptitude and L2 achievement than correlations between general intelligence and L2 achievement. It is important to note, however, that L2 aptitude and general intelligence are also correlated. In more general terms, L2 aptitude appears to be relatively independent of general intelligence, yet the two constructs also appear to overlap. Interestingly, this overlap seems to be centred around LAAs (Sasaki, 1996), and is another reason why more research into LAAs is warranted.
Another fundamental assumption of L2 aptitude is that it is relatively fixed or stable over the long term and resistant to training. The basis for this assumption seems driven more by theoretical speculation than empirical research, with little evidence in the literature to support such a claim. In fact, very little research in general has investigated this issue. This is surprising, given that real world observations suggest that learning a third language is achieved more easily than the second – an analogous observation to that underpinning the existence of L2 aptitude. What is more, a small body of research consistently supports this view that, contra the traditional assumption, L2 aptitude appears to improve with experience. More research is clearly needed into these two related issues of stability and trainability of L2 aptitude.

Stability and trainability also offer a clear opportunity for the practical application of L2 aptitude research. If experience leads to improvements in L2 aptitude and aspects of this experience are teachable, then research in the trainability of L2 aptitude potentially offers a way to enhance L2 learning outcomes for L2 learners and teachers. As a language teacher, the possibility of helping people to become better L2 learners was a strong motivator behind my current study.

Investigating the stability and trainability of L2 aptitude is also important for research in the field of SLA. If L2 aptitude is not stable, this would require explanations of L2 aptitude and its role in SLA to be updated. Conversely, if L2 aptitude was confirmed as stable and untrainable, then the research might offer insights and explanations as to why this is the case. In addition to these insights into L2 aptitude theory, research into L2 aptitude’s purported stability is also methodologically relevant to the measurement of L2 aptitude in SLA. For
example, if L2 aptitude is not stable and is, in fact, sensitive to experience, this may suggest that L2 aptitude testing needs to take place prior to L2 learning.

L2 aptitude has been researched for over 60 years, yet it has only recently reemerged as a topic of interest in SLA. This appears at odds with the fact that L2 aptitude tests are often considered the strongest ID predictor of L2 achievement (Sawyer & Ranta, 2001).

Historically, L2 aptitude research can be divided into three general periods. In the 1950s and 1960s, L2 aptitude research focussed on developing, validating, and standardising L2 aptitude tests (but see Henmon, 1929 for early work on prognosis testing of second languages). The most influential works of this period were Carroll and Sapon’s (1959) Modern Language Aptitude Test (MLAT) and Pimsleur’s (1966) Language Aptitude Battery (PLAB). The primary use of these tests was for selection – the MLAT for intensive L2 audiolingual training and the PLAB for high school L2 courses. However, once these tests were established, there seemed little more to be achieved in the field. Consequently, from about the 1970s through to the 1990s, research into L2 aptitude was limited. Although some important studies did take place during this time, they did not generate much interest in the field. Two possible explanations may account for this loss of interest. First, L2 aptitude tests were associated with the prevalent L2 teaching methodology of the time, most notably audiolingualism. As L2 teaching methodologies shifted towards more communicative-oriented methods, L2 aptitude tests and what they measured were seen to be less and less relevant. Second, L2 aptitude as a stable trait undermined the role of the learner in L2 learning. The view of L2 aptitude as immutable seemingly undervalued the importance of effort in L2 learning, so was not a widely popular concept with L2 educators. However, this view of L2 aptitude changed from the 1990s onwards. As SLA research interest shifted towards individual difference variables during the 1990s, L2 aptitude research increased.
Consequently, more recent L2 aptitude research has differed from that conducted in the past. In particular, the influence of the cognitive revolution in SLA research has impacted L2 aptitude research. For example, some of the more recent models attempting to explain L2 aptitude have been distinctly cognitive, from the application of Sternberg’s theory of triarchic intelligence in the CANAL-FT L2 aptitude test (Grigorenko, Sternberg, & Ehrman, 2000) to the adaptation of Snow’s (1991) theory of aptitude-treatment interactions in Robinson’s (2001, 2005) aptitude complex hypothesis. Developments in how L2 aptitude has been researched have also driven renewed interest in the area and offered valuable new insights. These new approaches include methodological techniques such as controlled experimental laboratory interventions, and new methods of statistical analysis such as structural equation modelling. Consequently, L2 aptitude research is rapidly advancing with new findings that have led to recent calls for the construct to be updated. However, it is important to note that, broadly speaking, the measurement of L2 aptitude has remained relatively unchanged for most L2 aptitude research, despite a number of scholars arguing the need for new and more targeted tests.

1.2 Theoretical framework
As mentioned above, more recent research into L2 aptitude has been cognitively based. This can be attributed to the cognitive basis of L2 aptitude tests as well as to the increase in cognitive-based explanations and research of the area. In line with these developments in the field of L2 aptitude research, the current study has also adopted a cognitive framework to investigate the issues of stability and its correlate, untrainability. In particular, the current study looked for explanations of L2 aptitude that could account for a dynamic construct that
was amenable to training. Clearly, the traditional model of L2 aptitude makes no allowances for such research and, therefore, offers no practical guidance as to how these related phenomena could be investigated. The cognitive models that did offer explanations of L2 aptitude as dynamic and amenable to training had common features. Each viewed L2 aptitude as based on general cognitive abilities organised in a hierarchical system involving L2 learners as adaptive agents in the L2 learning process, and regarded higher-order thinking involved in L2 learning as teachable. These explanations still operationalise L2 aptitude as test scores, but account for test performance in a more holistic and practical way. These holistic cognitive-based explanations were, therefore, adopted in the current study to explore the issues of stability and trainability in L2 aptitude.

1.3 Aims of the study

Despite the advances in L2 aptitude research, certain fundamental assumptions have remained relatively unquestioned. The stability of L2 aptitude is one of the least discussed and least researched issues in the literature on L2 aptitude. Although no conclusive evidence-based argument has yet been put forward to disprove the assumption of stability, the research suggests that L2 aptitude test scores are not stable and do improve with experience in L2 learning. Consequently, the primary focus of the current study is to investigate whether or not L2 aptitude is stable or sensitive to experience.

Related to the assumption of stability is that of untrainability. Evidently, if L2 aptitude is stable, then it follows that it is also untrainable. However, it does not follow that if L2 aptitude is not stable, then it must be trainable. The trainability in L2 aptitude has been even

\[\text{Carroll (1981) does state that he is sympathetic to a view of L2 aptitude as not fixed or innate, but that the (unpublished) evidence he has seen suggests that this is not the case.}\]
less studied than stability, with only two training studies having been conducted to date (Politzer & Weiss, 1969; Yeni-Komshian, 1965). Notably, no study explicitly investigating the trainability of L2 aptitude has been conducted in over 40 years. Consequently, the secondary focus of the current study is to explore whether or not L2 aptitude is trainable. This question is specifically investigated by concentrating on LAAs, one of the less understood components of L2 aptitude.

In investigating the issues of stability, and its correlate untrainability, in L2 aptitude, two additional and related gaps were identified in the research. First, to train LAAs requires a working model to develop an effective training intervention. As mentioned above, the literature has little to say on what actually constitutes language analytic abilities and empirical studies only offer indirect explanations. Second, as in most L2 aptitude research, the construct is operationalised by L2 aptitude tests. As was the case for LAAs, the literature on L2 aptitude tests does not offer concrete explanations or evidence for what the tests actually measure. For example, I could find no study that investigated individual’s behaviours while taking L2 aptitude tests to corroborate what were mainly inferred post-hoc descriptions of test tasks.

Investigating the stability and untrainability in L2 aptitude is a complex and challenging task, yet also offers extremely valuable insights into the construct of L2 aptitude. In particular, the current study seeks to expand the traditional explanations of L2 aptitude. For L2 aptitude to be recognised as a dynamic construct, it must be regarded as more than just four basic abilities. Inevitably, a model of L2 aptitude must account for the agency of the L2 learner and their abilities to plan, monitor, and evaluate the decisions they make and actions they take during their L2 learning. For L2 aptitude to be recognised as a dynamic construct, L2 learners
must also be viewed as adaptive and able to profit from the feedback they receive through
experience, regardless of whether that experience is ‘naturalistic’, instructed, or takes the
form of results from L2 aptitude tests. The traditional model of L2 aptitude does not allow for
such explanations, but cognitive models have been advanced that do. The current study seeks
to fill this gap in L2 aptitude research and add to our current understanding of the construct.

The aims of the current study are thus threefold. First and foremost, the current study
investigates whether L2 aptitude can improve over time. A second, related goal of the current
study is to investigate the trainability of L2 aptitude. In particular, the current study
investigates whether LAAs are trainable. If test scores for LAAs increase and it can be shown
that a training intervention was at least partially responsible for those increases, the claim that
L2 aptitude is untrainable in general is weakened. A third aim of the current study is to
contribute to the understanding of current L2 aptitude models. To investigate stability and
trainability, various problems first need to be solved. In particular, the development of a
LAAs training intervention involves a deeper understanding of what tests of LAAs actually
measure as well as how LAAs can be described. Again, these understandings can be
generalized to the L2 aptitude construct as a whole to help expand our current thinking about
both L2 aptitude measurement and explanation.

1.4 Overview of current study

The current study investigates the issues of stability and and its correlate, untrainability, in L2
aptitude by employing a pseudo-experimental field study design. By investigating L2 aptitude
with L2 learners in a real-world situation of university study, the current study seeks to
maximise its ecological validity. Consequently, the study investigates stability by pre- and
post-testing the L2 aptitude of L2 learners, which is done during the first 8-9 weeks of a beginner Spanish course. Untrainability is investigated by designing and implementing a 6-week training course in language analytic abilities, which participants completed between the pre- and post-tests. The training course was developed especially for the study and was strongly influenced by the cognitive accounts of L2 aptitude. Analyses included paired-samples $t$-tests to establish changes in L2 aptitude test scores and regression models to investigate the predictors of changes in post-test scores for LAAs.

1.5 Contribution to the field

Building on the previous research into the issues of stability and untrainability in L2 aptitude, the current study employs new theoretical, methodological, and analytic approaches to contribute to the field. Theoretically, the current study is the first training study in over 40 years and is the first to explicitly train LAAs. This undertaking also included proposing novel explanations for the construct of LAAs and how they relate to L2 learning. The two previous training studies attempted to train all aspects of L2 aptitude or only phonological abilities. Methodologically, the current study was the first training study to deliver a training intervention online, allowing participants to complete the training wherever and whenever they wished. This administration of the training allowed for more detailed data on participant behaviour to be captured as well as empowering participants to control their own learning, which was seen as increasing motivation in training performance. The two previous training studies administered their training sessions at designated times and places, regardless of participants’ state of readiness for instruction. In addition, the current study sampled from adult populations to minimise any maturational effects that may have confounded previous
studies of stability and untrainability in L2 aptitude. In terms of analysis, the current study was also the first to analyse effect sizes for longitudinal gains and was also the first to analyse these gains for sub-groups of below and above average L2 aptitude levels. Previous studies have not included effect sizes to quantify any gain scores that they found. Lastly, the current study was also the first training study to employ regression analyses to investigate not only if the training was predictive of gains, but also which aspects of the training were more important. Previous training studies only employed experimental vs control group comparisons as an analysis of the effectiveness of the training intervention. This comparison fails to account for whether L2 aptitude was untrainable or whether that particular type of training was ineffective.

The current study, therefore, seeks to contribute to the field of L2 aptitude research in terms of both theory and methodology. In respect to theory, if L2 aptitude tests are a proxy for L2 aptitude itself, then the current explanations of the construct need to be updated to take into account that L2 aptitude is sensitive to experience and may include teachable components. Such a view of L2 aptitude requires L2 aptitude tests to be reconceptualised as involving more than just measures of first-order abilities. Evidently, traditional L2 aptitude tests measure more than just phonological, memory, and language analytic abilities and might be more accurately regarded as measures of aptness as opposed to aptitude. Methodologically, studies of trainability in L2 aptitude need to move away from reliance on just experimental-control group comparisons and begin to investigate predictors of changes in L2 aptitude over time. Plainly, this raises many questions about the validity of L2 aptitude measured by a once-off static test, and suggests that alternative ways of measuring L2 aptitude should be considered, e.g. dynamic tests such as the CANAL-FT.
1.6 Outline of chapters

The above discussion contextualises, motivates, and reports on the thesis as a whole. The thesis is organised as follows.

Chapter 2 critically reviews the theory and research of L2 aptitude to answer the question: What is L2 aptitude? The discussion provides an overview of the traditional model of L2 aptitude and its limitations before reporting the results of a systematic review of L2 aptitude research. From this review, the issues of stability and its correlate, untrainability, in L2 aptitude emerge as important gaps in the research.

Chapter 3 extends the discussion and explores the theory and previous research on stability and untrainability in L2 aptitude assumptions in L2 aptitude. The basis for these assumptions is examined in terms of both theory and empirical research. In addition, cognitive models of L2 aptitude are discussed with a focus on how their expanded explanation of L2 aptitude is able to account for empirical challenges to the traditional model. The chapter then shifts its focus to the specific topic of the trainability of LAAs. Definitions and explanations are first examined before proposing a novel model for LAAs that forms the theoretical basis for the training intervention used in the study. The chapter concludes with the research questions of the current study.

Chapter 4 connects the theory of the previous two chapters and describes how the research questions are investigated. All decisions made in the design of the study are detailed and explained. The instruments in the study are also described, with a detailed explanation of each aspect of the training course provided.
Chapter 5 moves the discussion from the practical aspects of the study design to the analysis of the data collected. The statistical tests carried out for investigating all aspects of the study are given, and the entire process of the analysis is explained.

Chapter 6 contextualises the results from the previous chapter in the broader field of L2 aptitude research. Implications of the major findings of the study are also discussed in depth.

Chapter 7 concludes the thesis by revisiting the major findings and their implications as well as by identifying the limitations of the current study. The chapter and thesis conclude with a discussion of the way forward in L2 aptitude research in terms of both theory and research.
CHAPTER 2: WHAT IS SECOND LANGUAGE APTITUDE?

This chapter examines the theoretical and empirical research regarding second language (L2) aptitude. Section 2.1 critically reviews the definitions and theories of L2 aptitude and their ability to explain the research. In particular, the assumption of stability and untrainability is identified as an important issue that lacks empirical support (section 2.1.2). Alternative views of L2 aptitude are then discussed to identify the key issues of this assumption of stability and untrainability (section 2.1.3 to 2.1.5). Section 2.2 systematically reviews L2 aptitude empirical research to identify its generalisations, its limitations and/or biases, and areas that remain unclear and/or less researched. The results of this systematic review indicate a need for research into the assumption of stability and untrainability in L2 aptitude. How these results inform, contextualise, and motivate the design of the current study are also discussed.

2.1 Definitions and theories of L2 aptitude

L2 aptitude can be broadly described as a “talent” for language learning, although in terms of L2 aptitude research this talent is defined as the ease with and rate at which the L2 is initially required (Carroll, 1981). Used in this way, the term appears to refer to a real-world phenomenon, but fails to describe it in any objective or useful way. Definitions of L2 aptitude in the literature (cf. Dörnyei, 2005; Robinson, 2012; Sawyer & Ranta, 2001; Skehan, 1998, 2002) suffer from a lack of clarity, with no clear consensus emerging (Rogers, Meara, Barnett-Leigh, Curry, & Davie, 2017). Most variations in the definitions of L2 aptitude are based on Carroll’s work with the Modern Language Aptitude Test (MLAT) (Carroll & Sapon, 1959) or other instruments that seek to measure this “talent” for language learning (see Grañana, 2013 for the LLAMA, Grigorenko et al., 2000 for CANAL-FT, and Linck et al.,
Indeed, the only point of agreement among researchers would appear to be that L2 aptitude is what L2 aptitude tests measure (Dörnyei, 2005; Singleton, 2017). In other words, the theories of L2 aptitude are derived from the power of these tests to predict L2 achievement. As a result, much more work is needed to create a robust definition and theory of L2 aptitude (Singleton, 2017; Wen, Biedrón, & Skehan, 2017).

While L2 aptitude may not be readily defined, tests such as the MLAT have reliably predicted L2 achievement for over 50 years (Sasaki, 2012). Indeed, L2 aptitude is the single best predictor of L2 achievement along with L2 motivation (Dörnyei, 2005; Wen et al., 2017). Therefore, one way to better understand L2 aptitude is to examine both the development and use of these tests. In the following section, we will examine the MLAT and its development, as it remains the most validated and reliable measure of L2 aptitude to date. The remainder of the chapter will then systematically review studies that have incorporated explicit measures of L2 aptitude, in order to summarise and synthesise the state of current L2 aptitude research.

2.1.1 The MLAT as a proxy for L2 aptitude

Arguably, the Modern Language Aptitude Test (Carroll & Sapon, 1959) is the benchmark test of L2 aptitude, even after 50 years (Sasaki, 2012). Its benchmark status is due primarily to its high and consistent levels of validity ($r=.4-.6$; Sasaki, 2012) and reliability ($r=.55-.92$; ibid.) when applied to different modes of instruction as well as to naturalistic contexts (Sawyer & Ranta, 2001). The ease of administration – a paper-and-pencil test as well as a computer-based test that can be completed in 50-60 minutes (LLTF, 2013) – has also made it practical enough to withstand the test of time. As the breakthrough test for L2 aptitude (Spolsky, 1995), the MLAT set the stage for subsequent work in the field.
The MLAT was developed as a direct response to a United States Government need to “select a lot of government people who could easily learn foreign languages” (Stansfield & Reed 2004:48) in a cost effective and timely manner (Carroll, 1962). As such, its development was not driven by theory but instead by more pragmatic needs. Carroll (1962) explains how the test was developed through a process of trial and error, typical of early psychometric test development (Spolsky, 1995). Subtests of L2 aptitude were trialled, analysed through factor-analysis, and then correlated, until Carroll and Sapon eliminated tests with little predictive value and combined tests that replicated each other (Carroll, 1962). What resulted became the MLAT (Dörnyei & Skehan, 2003). This process was chosen for its resulting high levels of test validity and reliability (Stansfield & Reed, 2004). Indeed, the MLAT has been adopted by all English-speaking governments (C.W. Stansfield, personal communication, September 28, 2013). Testament perhaps to the test's success is the fact that it remains unchanged since its initial publication, with only one version available (Carroll, 1990; Sasaki, 2012). However, multiple versions would have been useful for pre- and post-testing individuals, increasing the integrity of the test (from leaks), and investigating the effects of instruction, and indeed of different methods of instruction, in L2 learning (Carroll, 1990).

The development of the MLAT was based upon two novel and important conceptualisations of L2 aptitude. First, Carroll (1962) conceived of L2 aptitude as language specific – not in a theoretical sense but in terms of operationalisation. He clearly identified the failure of tests of general intelligence to reliably predict L2 achievement as evidence that aptitude for language learning was independent of general intelligence, although the MLAT IV Words in Sentences test did correlate with general intelligence measures (Carroll, 1962). So, Carroll focussed on tests that implicated abilities and processes logically involved in L2 acquisition, such as remembering vocabulary and decoding phonemic systems. While this assumption led Carroll
to construct the first and valid measure of L2 aptitude (Spolsky, 1995), there is evidence that L2 aptitude may not be as independent of general intelligence as Carroll originally assumed (Sasaki, 1993). Second, Carroll conceived of L2 aptitude not as the capability to acquire another language, but rather as the speed and facility with which individuals are able to learn other languages. Again, this distinction was not theory-based, but emerged from strictly pragmatic considerations. The objective of the MLAT was to identify individuals worthy of government and military investment for intensive language study. Despite its importance in SLA theory, the focus of inquiry was on the accurate prediction of an individual’s ability to succeed according to the criteria of a particular language course, rather than on predicting what level of L2 proficiency an individual might ultimately attain.

In summary, Carroll (1962) delimited his construct of L2 aptitude to a subset of cognitive and perceptual linguistic abilities that lead to faster and easier language learning. For the purposes of the MLAT, learning another language involved intensive and full-time formal instruction for approximately 1-2 years with specified criteria defining high levels of L2 achievement.

One final point worth noting about the MLAT from a practical perspective is the use of work-sample tests. Carroll (1962:97) notes in this regard that:

> As a sample of the kinds of tasks to be learned, [the work-sample test] may very well require the same abilities that are required in a broad class of criterion tasks. The abilities may appear to be specific to this class of criterion tasks only because nobody has invented a way of tapping them in other connections. (Carroll 1962: 97)

Essentially, Carroll is claiming that the validity of the MLAT tests to reliably predict L2 achievement indicates that the abilities tested in these work-sample tests are implicated in L2
acquisition as a whole. Spolsky (1995:335) succinctly summarises Carroll’s approach, noting that:

If he could, he wanted to find tests that tapped the most basic abilities in language learning, the discrete primary skills. Failing this, he sought to find the smallest trial learning situation that would predict performance in a full course.

Two important points in relation to this approach are to be borne in mind when considering Carroll’s theoretical work on L2 aptitude. First, Carroll’s use of work-sample tasks in aptitude lacks theoretical rigour and instead relies on statistical predictive power. Therefore, any subsequent theoretical claims Carroll (1962, 1973, 1981, 1990) makes about L2 aptitude are subject to the limitations of the use of work-sample tests. In other words, the theory derived from the MLAT is speculative. Second, and correlatively, while Carroll claims that the criticism of work-sample tests as overly specific is unfounded, arguing that only a few tests are needed to satisfactorily measure L2 aptitude, he fails to account for what the work-sample tests actually measure. For example, most of the MLAT test tasks are of a problem-solving type. Problem-solving is goal oriented (Reed, 2000), sometimes with many ways of achieving any particular goal (Galotti, 2013). Thus, the problem with work-sample tests is that although all individuals may confront a problem with the same set of abilities, they are free to strategically employ those abilities as they see fit (Sasaki 1993b). Thus, the work-sample tasks of the MLAT are impure measures of L2 aptitude abilities and any post-hoc theory derived from these tasks is questionable without empirical supporting evidence. Consequently, it remains unclear exactly what abilities constitute L2 aptitude.

Let us now turn to the theoretical conclusions Carroll derived from the MLAT, where the above issues are particularly salient.
2.1.2 Carroll’s four-factor construct of L2 aptitude

Once the MLAT was fully developed and standardised, Carroll (1962) inferred a post-hoc formulation of L2 aptitude by analysing the tests with the aid of results from factor analysis. This resulted in Carroll’s four-factor construct of L2 aptitude (see table 2.1), which continues to be extremely influential to this day (Skehan, 2002; Wen et al., 2017).

Table 2.1 Carroll’s (1962:129-130) four factors of L2 aptitude

<table>
<thead>
<tr>
<th>Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonetic Coding</td>
<td>the ability to meaningfully store auditory information for access at a later time</td>
</tr>
<tr>
<td>Grammatical Sensitivity</td>
<td>the ability to “handle grammar” and discern the functions of words in various contexts</td>
</tr>
<tr>
<td>Rote Memory for Foreign Language Materials</td>
<td>the ability and capacity to memorise a number of associations from the input materials, similar but independent to phonetic coding, encompassing more than just phonetic information</td>
</tr>
<tr>
<td>Inductive Language Learning</td>
<td>“the ability to infer linguistic forms, rules, and patterns from new linguistic content...with a minimum of supervision or guidance”</td>
</tr>
</tbody>
</table>

As this theoretical construct is derived from the MLAT, it can be argued that to understand L2 aptitude one needs to understand the test itself. Table 2.2 describes each sub-test of the MLAT and which of the four factors each sub-test purportedly measures (example questions from each test can be found at http://lltf.net/mlat-sample-items).
Table 2.2 The sub-tests of the MLAT and the factors they purportedly measure (Carroll, 1962; Sasaki, 2012)

<table>
<thead>
<tr>
<th>Sub-Test</th>
<th>Description</th>
<th>Main Factor Tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part I: Number Learning</td>
<td>Candidates learn numbers in an unknown language and then translate new numbers from this language into English.</td>
<td>Rote Memory for Language Learning Materials</td>
</tr>
<tr>
<td>Part II: Phonetic Script</td>
<td>Candidates learn associations between phonetic sounds and symbols before choosing the correct symbol for a new sound.</td>
<td>Phonetic Coding</td>
</tr>
<tr>
<td>Part III: Spelling Clues</td>
<td>Candidates choose one of five words with the closest meaning to a given word in a reduced spelling format.</td>
<td>Phonetic Coding</td>
</tr>
<tr>
<td>Part IV: Words in Sentences</td>
<td>Candidates identify the function of words in different sentences.</td>
<td>Grammatical Sensitivity</td>
</tr>
<tr>
<td>Part V: Paired Associates</td>
<td>Candidates learn the English meanings for a set of words in an unknown language and then afterwards complete a multiple choice quiz on the same word set.</td>
<td>Rote Memory for Foreign Language Materials</td>
</tr>
</tbody>
</table>

One obvious problem here is that one of Carroll’s four factors of L2 aptitude, the inductive language learning ability, is not identified as the main factor tested by any of his MLAT subtests (Carroll, 1962). How, then, does a factor analysis of the MLAT result in an aptitude ability that is not measured by the test? This highlights the speculative nature of Carroll’s construct and its being based, at least in part, on Carroll’s intuitions and experience. A theory of L2 aptitude is surely weakened by a lack of clear supporting empirical evidence.
What, then, do the work-sample test tasks of the MLAT actually measure? While Carroll’s construct suggests that the sub-tests are relatively pure measures of three of the four aptitude factors, no evidence definitively corroborates this claim. Surprisingly, perhaps, given the amount of criticism the MLAT has received for its atheoretical basis (see Sáfár & Kormos, 2008; Dörnyei & Skehan, 2003; Dörnyei, 2005; and Sawyer & Ranta, 2001 for detailed reviews of criticisms), few reanalyses of the test tasks themselves have been conducted (see Skehan, 2002 for discussion). A closer inspection of the test tasks for the factors they actually measure, gives rise to alternative perspectives. Take the MLAT IV Words in Sentences, for example.

The MLAT IV is a test of recognition and analogy of the grammatical function of words in sentences (Carroll, 1962). An example question is given here in figure 2.1, taken from the LLTF website (LLTF, 2013).

In each of the following questions, we will call the first sentence the key sentence. One word in the key sentence will be underlined and printed in capital letters. Your task is to select the letter of the word in the second sentence that plays the same role in that sentence as the underlined word in the key sentence.

Look at the following sample question:

Sample: **JOHN** took a long walk in the woods.

| Children in blue **jeans** were **singing** and **dancing** in the **park**. |
| A | B | C | D | E |

*Figure 2.1 Example question from the MLAT IV Words in Sentences sub-test*

Carroll claims that this test measures “the ability to recognise the grammatical function of words in sentences... without... requiring the examinee to know the meaning of... grammatical terms” (Carroll, 1962:95). However, the test most likely measures other factors besides just this recognition and analogising of the grammatical functions of words. For one thing, it
measures the L1 ability to parse and analyse sentences into their constituent parts, as all questions are in English for L1 speakers. To this end, the test must measure L1 literacy skills (Sáfár & Kormos, 2008). For another, it must also, inevitably, measure some metacognitive skills, e.g., test-wiseness. In other words, to be able to answer this question, the examinee must first comprehend the sentences themselves, which means being able to read them. This relies on L1 phonological abilities to be able to decode the words and then construct the sentence with its constituent parts (Sparks, Patton, Ganschow, & Humbach, 2009). However, this still leaves the examinee with an ambiguity in relation to the key words. ‘JOHN’ can be analysed as a noun, a noun phrase, subject, or agent. But knowing this allows the examinee to at least eliminate singing (C) and dancing (D) from a list of possible answers – a common strategy for answering multiple choice questions. Without taking into account any explicit (metalinguistic) knowledge of grammar that may be brought to bear, even as brief an analysis as this of what may be involved in completing the MLAT IV test task, suggests that a range of factors beyond grammatical sensitivity may be implicated.

Put simply, such a summary analysis demonstrates that Carroll’s post-hoc construct of L2 aptitude derived from the MLAT warrants further research. In particular, Carroll’s assumptions that are not supported by empirical evidence require more investigation, such as the assumption of specificity (Sasaki 1993a) and the assumption of stability (Carroll, 1981; Robinson, 2012; Skehan, 2002). The current study focuses on the latter, as this is the less researched and less understood of the two (see section 2.2.5.5 and 2.2.5.8 for comparisons). However, despite these criticisms, Carroll’s contribution to L2 aptitude research is immeasurable. Predictive L2 aptitude tests such as the Pimsleur Language Aptitude Battery (PLAB) (Pimsleur, 1966), the Defence Language Aptitude Battery (DLAB) (Petersen &
Al-Haik, 1976), and the VORD (Parry & Childs, 1990) adopted the same essential construct as Carroll’s MLAT (Kiss & Nikolov, 2005), with slight differences in how the tests were operationalised. More recently, the LLAMA aptitude tests, which are loosely based on the MLAT (Meara, 2005), have been increasingly used in L2 aptitude research (Granena, 2013; Rogers et al., 2017). Specific training tasks in the current study were also highly influenced by Carroll’s tests and ideas of grammatical sensitivity and inductive language learning in particular (see section 4.4.3). As mentioned, the MLAT was the first evidence for L2 aptitude (Carroll, 1962; Spolsky, 1995). The factor analyses of L2 aptitude measures clearly showed the construct to be componential and not unitary (Carroll, 1981). Further work based on the components of L2 aptitude has also shown that L2 aptitude can be broadly categorised into phonological abilities, language analytic abilities, and memory abilities (Sasaki, 1993a, 1993b; Skehan, 2002). Finally, the fact that the MLAT reliably predicts L2 achievement across a range of contexts gives us insight into second language learning, even if it remains unclear exactly what the test measures. Nevertheless, the shortcomings of early and traditional L2 aptitude theory have led to various alternatives being proposed. In the following sections, we will briefly discuss these alternative theories and identify some of the issues still to be resolved.

2.1.3 Skehan’s (2002) model of L2 aptitude and SLA

Skehan’s (2002) hypothetical model of L2 aptitude attempts to match the major processes of SLA with implicated L2 aptitude abilities, creating an integrated model of the two constructs. Table 2.3 shows the simplicity of Skehan’s integrated model, with the information processing stages of SLA on the left and the corresponding factors of L2 aptitude implicated in those
processes on the right. What results is a model with both testable and explanatory hypotheses (Robinson, 2012). For example, an L2 learner’s ability to identify syntactic patterns from input could be predicted by existing tests that measure working memory capacity (e.g. a reading span test) and grammatical sensitivity (e.g. MLAT IV-Words in Sentences test, see table 2.2). These measures of aptitude could then be correlated with defined and measurable levels of achievement of identified syntactic patterns (such as through a grammaticality judgement test).

Table 2.3 *SLA processing stages and potential aptitude components (Skehan, 2002:90)*

<table>
<thead>
<tr>
<th>SLA Processing Stage</th>
<th>Corresponding aptitude constructs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Noticing</td>
<td>Auditory segmentation</td>
</tr>
<tr>
<td></td>
<td>Attention management</td>
</tr>
<tr>
<td></td>
<td>Working Memory</td>
</tr>
<tr>
<td></td>
<td>Phonemic Coding</td>
</tr>
<tr>
<td>2. Pattern Identification</td>
<td>Fast Analysis/Working Memory</td>
</tr>
<tr>
<td></td>
<td>Grammatical Sensitivity</td>
</tr>
<tr>
<td>3. Extending</td>
<td>Inductive Language Learning Ability</td>
</tr>
<tr>
<td>4. Complexifying</td>
<td>Grammatical Sensitivity</td>
</tr>
<tr>
<td></td>
<td>Inductive Language Learning Ability</td>
</tr>
<tr>
<td>5. Integrating</td>
<td>Restructuring Capacity</td>
</tr>
<tr>
<td>6. Becoming Accurate, Avoiding Error</td>
<td>Automatisation</td>
</tr>
<tr>
<td></td>
<td>Proceduralisation</td>
</tr>
<tr>
<td>7. Creating a Repertoire, Achieving Salience</td>
<td>Retrieval Processes</td>
</tr>
<tr>
<td>8. Automatising Rule-based Language, Achieving Fluency</td>
<td>Automatising, Proceduralisation</td>
</tr>
<tr>
<td>9. Lexicalising, Dual-coding</td>
<td>Memory, Chunking, Retrieval Processes</td>
</tr>
</tbody>
</table>

Skehan's model shows promise in two main areas: as an agenda for future research and guidance for the development of new instruments. By matching SLA processes with L2 aptitude components, and by further organising the nine SLA stages into four macro-stages –
noticing (stage 1), patterning (stages 2-5), controlling (stages 6-8), and lexicalising (stages 8-9) – new predictions of how L2 aptitude components will affect SLA processes emerge. For example, noticing will be affected by phonetic coding and patterning will be affected by both grammatical sensitivity and inductive language learning. Skehan also integrates into this model of L2 aptitude more modern components, such as working memory, as well as components not previously posited, e.g. restructuring capacity. But most importantly, all these L2 aptitude constructs can be tested within the context of their corresponding SLA process (see Robinson, 2012 for a description of how this could be done).

As a result of Skehan’s (2002) model, in which L2 aptitude components are matched to a context in SLA, new instruments for measuring L2 aptitude can be imagined. For example, fast analysis could easily be derived from existing measures, such as the MLAT IV or the PLAB IV, where the task is based on pattern identification. All that would be needed would be to add a time dimension to the task, whether that be restricting overall time or recording how long each question takes to complete. These considerations were key in the development of problem solving tasks in the training course created as part of the current study (see section 4.4.3.2). This example of modifying existing measures shows how Skehan’s model could find new lines along which to measure and discriminate L2 aptitude abilities in individuals.

However, the model proposed by Skehan is limited. In making the case for matching SLA processes with L2 aptitude abilities, Skehan (2002) himself admits that the model of SLA he utilises is not accepted unanimously. Much more work is needed both theoretically and empirically in both these models of SLA and L2 aptitude.
A further shortcoming is that Skehan suggests neither how nor when all abilities could or should be measured. For example, automatising and proceduralisation are both proposed as L2 aptitude components at the controlling stage of SLA. However, automatising is a long-term process (McLaughlin, 1995) and relies upon information already learned. Presumably, testing automatising with novel linguistic information would then prove difficult. The other problem concerns the time frame over which automatising would be tested. The MLAT and PLAB are one hour at the most – would this be long enough to test automatising? Or would automatising need to be tested at a later stage? How would that kind of testing be practical? These are just some of the questions and challenges confronting the measurement of the novel L2 aptitude abilities proposed by Skehan (2002).

Another limitation is the fact that many of the abilities Skehan proposes to test, particularly in the latter two stages, are productive skills. Almost no work has been done in L2 aptitude testing in this respect, making this area particularly difficult to investigate and potentially beyond the scope of current L2 aptitude research.

In summary, Skehan has proposed an ambitious theoretical framework and research agenda in his model of L2 aptitude in SLA. It attempts to update current constructs of L2 aptitude by incorporating the developments of the last 40 years since Carroll first proposed his four factor construct. With testable hypotheses and a theoretical base, it will be interesting to see how well the model works. Skehan also lays down significant challenges for future research with a more nuanced view of L2 aptitude and its potential fit in the big picture of SLA.
2.1.4 Robinson’s Aptitude Complex Hypothesis

Like Skehan, Robinson’s (2001, 2005) Aptitude Complex Hypothesis attempts to describe L2 aptitude in great detail. However, Robinson takes a radically different view. Rather than viewing L2 aptitude as a single construct applied to the single task of learning a L2, Robinson’s is a more detailed perspective of both L2 aptitude and SLA. Robinson (2001, 2005) proposes that certain abilities of L2 aptitude are drawn upon depending on the demands of specific L2 learning tasks. For example, learning grammatical patterns from input will draw upon the ‘metalinguistic rule rehearsal’ (MRR) aptitude abilities of grammatical sensitivity and rote memory, but will not necessarily draw upon the aptitude abilities of ‘memory for contingent speech’ (MCS), phonological working memory capacity, or phonological working memory speed. From this perspective of aptitude-task interaction (ATI), L2 aptitude becomes dynamic in its application, as the learner draws on the abilities needed to complete a given task, giving rise to various complexes of aptitudes determined by the demands of the task at hand.

Figure 2.2 shows how lower-level general cognitive abilities build complexes of aptitudes that play out in various learning tasks or contexts, resulting in pragmatic abilities or traits (Dörnyei, 2005). The figure illustrates the hierarchical structure of L2 aptitude. The first circle describes the cognitive abilities that underpin the various possible areas of aptitude (the second circle), which relate to the more traditional, Carrollian view of L2 aptitude. The third circle relates to L2 learning tasks and its characteristics, while the fourth relates to pragmatic performance (see table 2.4 for details).

Two points are worth noting here that distinguish Robinson’s hypothesis from the more traditional construct of L2 aptitude. First, some of the first order abilities here are novel,
especially the distinction between phonological working memory capacity (PWMC) and phonological working memory speed (PWMS), a distinction Robinson also makes for text working memory, presumably a reference to the visuospatial sketchpad of WM – a visual subsystem of working memory (Baddeley, 2003) – as applied to language learning. Second, while the second order aptitude complexes seem to be the equivalent of factors of L2 aptitude as a whole, Robinson has speculatively reconceived these last from phonological, memory, and language analytic abilities into noticing the gap, memory for contingent speech, deep semantic processing, memory for contingent text, and metalinguistic rule rehearsal. This results in a much more fine-grained description of L2 aptitude than Carroll’s traditional construct.
Figure 2.2 Aptitudes, development, and learning contexts: Changes in the relative contribution of aptitude factors to different aspects of L2 learning. (Inner two circles: initial input-based learning; third circle: output practice and complex task performance; and outer circle: transfer of task performance to real world interactive settings.)

(Robinson, 2005: 52)
Table 2.4 Key to Figure 2.2 (Robinson 2005:53)

<table>
<thead>
<tr>
<th>Abilities (inner circle)</th>
<th>PS = Processing Speed; PR = Pattern recognition; PWMC = Phonological Working Memory Capacity; PWMS = Phonological Working Memory Speed; SP = Semantic Priming; IN = Lexical Inferencing; TWMC = Text Working Memory Capacity; TWMS = Text Working Memory Speed; GS = Grammatical Sensitivity; RM = Rote Memory Aptitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexes (second circle)</td>
<td>NTG = Noticing the Gap; MCS = Memory for Contingent Speech; DSP = Deep Semantic Processing; MCT = Memory for Contingent Text; MRR = Metalinguistic Rule Rehearsal</td>
</tr>
<tr>
<td>Task Aptitudes (third circle)</td>
<td>+/- ST = Single Task; +/- PT = Planning Time; +/- BK = Background Knowledge; +/- H&amp;N = Here-and-Now; +/- FE = Few Elements; +/- R = Reasoning; +/- O = Open Task; +/- 1way = 1-Way Task; +/- CON = Convergent Task; +/- SG = Same Gender Participants; +/- SP = Same Proficiency Participants; +/- FAM = Familiar Participants</td>
</tr>
<tr>
<td>Pragmatic/Interactional Abilities/Traits (fourth circle)</td>
<td>II = Interactional Intelligence (Levinson, 1995); SP/IM = Self Presentation/Impression Management (Goffman, 1967); MR = Mind Reading (Baron-Cohen, 1995); PA (NLSC) = Pragmatic Ability (Nonliteral speech comprehension; Langdon et al., 2002); SI (SIT) = Social Insight (Social Insight Test; Chapin, 1967); EI (MEIS) = Emotional Intelligence (Multifactor Emotional Intelligence Scale; Mayer et al., 2000); SE = Self-Efficacy (Bandura, 1986); OTE (NEO) = Openness to Experience (Neuroticism, Extroversion, Openness Personality Inventory; Costa &amp; MacRae, 1985); GR = Gesture Reading (Goldin-Meadow et al., 1993); NVS (PONS; SIT) = Nonverbal Sensitivity (Profile of Nonverbal Sensitivity Test; Social Interpretation Test; Rosenthal et al., 1979; Archer, 1983)</td>
</tr>
</tbody>
</table>

Derived from Snow’s interactionist perspective and aptitude complexes (1994), the Aptitude Complex Hypothesis is an ambitious attempt to encompass, from input to output, the interaction between the “inner environment” (Snow & Lohman, 1989:347) of the learner and the “outer environment” (ibid.) of the learning context that results in SLA. In this view, aptitude is not internal to the learner but emerges in the interaction between the learner’s abilities and the learning situation (Sternberg, 2002). Consequently, Robinson’s research program has been concerned with individual differences and language instructional treatments (see Robinson, 1996, 1997, 2002a).

The Aptitude Complex Hypothesis still needs much work to demonstrate its relevance and accuracy. First, the hypothesis may be too restrictive in its detail and specificity. For example, the aptitude complex of metalinguistic rule rehearsal is dependent upon input, therefore phonological abilities may need to be taken into account, as there is evidence that these abilities underpin reading skills (Sparks, 1995). Inductive language learning is also seemingly ignored here or reimagined as something simpler, such as rote memory. As it currently stands,
this construct fails to take these factors into account. Second, Robinson, much like Skehan, proposes new abilities for which no instruments currently exist. This makes the Aptitude Complex Hypothesis untestable to an extent and is indicative of the work that remains to be done.

However, Robinson (2001, 2005) offers new insight into L2 aptitude informed by theoretical developments in both psychology and SLA (see Robinson, 2012). The Aptitude Complex Hypothesis adds detail and depth to the theoretical description of how L2 aptitude operates in the context of SLA, making testable predictions. In incorporating L2 aptitude and SLA theory, Robinson makes L2 aptitude more relevant to current research into language learning as well as offering a theoretical base for future research. Indeed, two ideas Robinson proposes influenced the design of test tasks and the measurement of training performance in the current study. First, Robinson represents language analytic abilities in his model in the form of ‘metalinguistic rule rehearsal’ (MRR), which seems to suggest that metalinguistic knowledge plays a role in language analysis. In fact, scores on the MLAT IV, a measure of grammatical sensitivity, seem to be influenced by metalinguistic knowledge (Carroll, 1962). Metalinguistic knowledge involves explicit knowledge of language, including the rules (Alderson, Clapham, & Steel, 1997), which suggests language analytic abilities also involve deductive reasoning from rules to examples. As a result, language analytic abilities in this study were conceived of as including deductive reasoning (see section 3.2.4), and tasks involving deductive reasoning were created and included in the treatment (see section 4.4.3.5). Second, Robinson explicitly names processing speed as a component of L2 aptitude. Like Skehan, Robinson suggests that the speed with which material can be processed is an indication of facility in L2 learning. L2 aptitude measures based on the MLAT do not currently take this factor into account. Consequent to Robinson, processing speed was
included as a variable of interest and operationalised as the time taken to complete L2 aptitude tests and treatment training tasks (see section 5.2.1).

While much work remains to be done, Robinson’s (2001, 2005) Aptitude Complex Hypothesis offers a detailed theory of L2 aptitude with powerful new insights and detailed glossing of the mechanics of L2 aptitude.

2.1.5 Cognitive Ability for Novelty in Acquisition of Language (Foreign) Test (CANAL-FT) Construct

Since the MLAT and Carroll’s (1962) original construct of L2 aptitude, paradigms have shifted in the fields of SLA, psycholinguistics and language learning (Carroll, 1990; Sáfár & Kormos, 2008; Sasaki, 2012). This led Grigorenko, Sternberg, and Ehrman (2000) to conceptualise L2 aptitude testing with their Cognitive Ability for Novelty in Acquisition of Language (Foreign) Test (CANAL-FT). Like the MLAT, the CANAL-FT takes a componential view of L2 aptitude. Unlike the MLAT, however, the CANAL-FT conceptualises language learning more generally, understanding it to be built primarily upon the ability to deal with novelty and ambiguity (Grigorenko et al., 2000), one of the components of Sternberg’s triarchic theory of successful intelligence (Sternberg, 2003).

At first glance, this may seem to construe L2 aptitude as simply a subset of intelligence. However, it must be borne in mind that Sternberg’s (2003) theory of successful intelligence is much broader than traditional concepts. Successful intelligence is based on the idea that individuals adapt to the demands of a task, taking advantage of their strengths and compensating for their weaknesses, drawing from a whole range of basic cognitive abilities (Sternberg, 2002) reminiscent of Robinson’s (2001, 2005) Aptitude Complex Hypothesis.
Consequently, from Sternberg’s perspective, L2 aptitude describes a subset of general cognitive abilities drawn upon to learn another language. As part of the broader theory of successful intelligence, language learning is just another task, albeit one with a unique set of demands that differ from those of other types of tasks, e.g. driving a car.

The CANAL-FT differs from Carroll’s traditional L2 aptitude construct in two key ways: it is grounded in cognitive theory and it takes a dynamic, rather than a static, view of L2 aptitude (Grigorenko et al., 2000). On this construal, L2 aptitude is dynamic, sensitive to experience, and, therefore, trainable (see section 3.1 for discussion).

Table 2.5 describes the five types of knowledge acquisition processes for learning in general that are applied to SLA posited and operationalised by the CANAL-FT (Grigorenko et al., 2000:392-393). The CANAL-FT is a simulation test of L2 learning which measures the learning as it occurs. Over nine test sections, test-takers are gradually exposed to an artificial L2, called Ursulu. The test measures the acquisition of certain lexical, morphological, semantic, and syntactic aspects of Ursulu so that by the end, the test-taker can process a small story in the artificial L2. Learning takes place in both visual and auditory modes and is measured in both the short- and long-term through immediate and delayed recall, respectively. The five learning processes hypothesised to underpin L2 learning are integrated across five different test tasks, shown in table 2.6. While the descriptions are informative, they do not clearly explain whether these processes reflect conscious and/or unconscious processes.
Table 2.5 CANAL-FT acquisition processes

<table>
<thead>
<tr>
<th>Knowledge Acquisition Process</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selective Encoding</td>
<td>The process whereby the relevance of incoming information is decided.</td>
</tr>
<tr>
<td>Accidental Encoding</td>
<td>The process of background or secondary information giving context to the information being focussed on.</td>
</tr>
<tr>
<td>Selective Comparison</td>
<td>The process of deciding the relevance of old information held in working memory on the current incoming information.</td>
</tr>
<tr>
<td>Selective Transfer</td>
<td>The process of applying previously decoded or inferred rules to the current situation or context.</td>
</tr>
<tr>
<td>Selective Combination</td>
<td>The process of synthesizing all information from selective and accidental encoding with old information structures, then incorporating these into a new and plausible unit of knowledge.</td>
</tr>
</tbody>
</table>

The CANAL-FT construct differs most noticeably from those discussed above by the way in which its description of language learning emphasises both the building of a system of knowledge, and the abilities of memory and recall involved. Of particular relevance to the current study, the building of this system of knowledge is dynamic and is thus influenced by pre-existing knowledge. On this understanding, language learning and thus L2 aptitude are dependent upon prior experience (Grigorenko et al., 2000). Consequently, L2 aptitude as a form of information processing, like other forms of information processing, can be developed (Sternberg, 1998). The CANAL-FT construct thus provides the theoretical and evidential support for the dynamism of L2 aptitude (Grigorenko et al., 2000), a consideration which will be discussed in Chapter 3. Sternberg (2002) also argues that both analytical and practical skills can be taught, suggesting that appropriately targeted L2 aptitude training could be
effective. This view of L2 aptitude as sensitive to previous experience\(^2\) and, therefore, trainable is key to the current study, which creates a theoretical and empirical base from which to question the assumption of stability and untrainability of L2 aptitude (see section 3.1 for detailed discussion).

Table 2.6 CANAL-FT test tasks and their cognitive processes and levels of processing

<table>
<thead>
<tr>
<th>Test task</th>
<th>Cognitive processes</th>
<th>Levels of processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning meanings of neologisms from context</td>
<td>selective encoding, accidental encoding, selective comparison, selective transfer, selective combination</td>
<td>lexical, morphological, semantic</td>
</tr>
<tr>
<td>Understanding the meaning of passages</td>
<td>selective encoding, accidental encoding, selective comparison, selective combination</td>
<td>semantic</td>
</tr>
<tr>
<td>Continuous paired-associate learning</td>
<td>selective encoding, selective comparison, selective combination</td>
<td>lexical, morphological</td>
</tr>
<tr>
<td>Sentential inference</td>
<td>selective encoding, accidental encoding, selective comparison, selective transfer, selective combination</td>
<td>morphological, syntactic</td>
</tr>
<tr>
<td>Learning language rules</td>
<td>selective encoding, accidental encoding, selective comparison, selective combination</td>
<td>lexical, morphological, semantic, syntactic</td>
</tr>
</tbody>
</table>

\(^2\) The term “experience” is used in thesis to cover both non-training exposure (e.g. classroom L2 instruction) and training treatment regimes.
2.1.6 L2 aptitude theory: Conclusions

In summary, Carroll’s (1962) original four-factor construct of L2 aptitude created an important base from which further research and theory has been built. The evolution of this construct delimited the groups of abilities understood to constitute L2 aptitude, namely phonological, memory, and language analytic abilities (Skehan, 2002). However, only the more recent constructs of L2 aptitude have been able to explain with any detail the processes that underpin L2 aptitude and integrate this construct with wider SLA theory (Skehan, 2002; Robinson, 2001, 2005) and learning in general (Grigorenko et al., 2000; Sternberg, 2002).

Importantly, new conceptualisations of L2 aptitude offer more insight into the question of its stability. These conceptualisations have informed the design of materials appropriate for the training of underlying abilities (see section 4.4.3).

However, these theories only pertain to one aspect of L2 aptitude. The last 60 years or so has seen much research into L2 aptitude, even though this has waxed and waned (Dörnyei, 2005; Sawyer & Ranta, 2001; Skehan, 2002). We now turn to this research and examine the results of a systematic review that sheds light on what else can be learned from over 50 years of L2 aptitude study.

2.2 Systematic review of L2 aptitude research

In the last 58 years, research into L2 aptitude has been infrequent, possibly due to the generalised traditional belief that L2 aptitude is fixed, and that L2 achievement is therefore predetermined, irrespective of the effort put into L2 learning (Skehan, 2002). Although new ideas have emerged, many reviews of L2 aptitude still focus on traditional questions,
critiques, and assumptions (see Dörnyei, 2005; Robinson, 2012; Sawyer & Ranta, 2001; Skehan, 1998, 2002). Only two meta-analyses have been conducted (Li, 2014, 2015), which despite making valuable contributions to the field, are limited to correlational data. Hence the pressing need for an objective and comprehensive review of L2 aptitude research. A systematic review of L2 aptitude research that goes beyond correlational data can uncover broader generalisations and limitations of the field, as well as identify areas that remain insufficiently researched. Consequently, this systematic review has three principle objectives:

1. To identify what generalisations can be drawn from the findings of the research.
2. To identify limitations and possible biases in the empirical evidence.
3. To identify areas in the L2 aptitude research that remain unclear or less researched.

These objectives relate directly to the current study. First, the generalizations of the findings of the research into L2 aptitude are used to ensure that the current study is situated within the current paradigm of research and that it is comparable to published studies. For example, one of the reasons the current study selects participants from the university population is because this is the most common group of people studied for L2 aptitude (see section 2.2.4.4). Second, identifying biases and limitations as well as less researched issues in L2 aptitude is used to ensure that the current study makes a valuable contribution to the field. For example, stability and untrainability are established as two of the least researched issues in L2 aptitude (see table 2.7), despite their being fundamental assumptions of the construct (see section 2.1.2). In short, the results of the systematic review are key in situating the current study and maximising its contribution to the field.

The review is structured as follows. First, the methods for designing and conducting the systematic review are detailed to ensure its objectivity and reproducibility (sections
2.2.1-2.2.3). Second, the results are presented, summarising the general characteristics of L2 aptitude research, discussing its findings, and identifying the areas in need of further investigation (sections 2.2.4-2.2.5).

2.2.1 Procedure for systematic review

The current systematic review was conducted following the guidelines set out in Pickering and Byrne (2014). To be included in the review, each paper had to meet all four of the following criteria: (1) be a journal article, PhD thesis, or a report; (2) pertain to empirical research that included some aspect of L2 aptitude; (3) utilise an instrument explicitly designed to measure L2 aptitude or one of its components; and (4) be published in English. Book chapters were excluded as a resource, because ensuring a comprehensive and exhaustive search of all the empirical studies available was not practically possible.

Scholarly electronic databases were searched to find empirical studies. These databases included: ERIC, JSTOR, ProQuest, SAGE, Web of Science, Wiley. These databases were searched between 07 May 2016 and 28 November 2016. For each database, an iterative keyword search was conducted. For details of the keywords and search procedure see Appendix A.

Following the guidelines in Pickering and Byrne (2014), the following data categories were created: publication details, geographical information, research methods used, participant data, variables measured, the discipline of each study, and findings and conclusions.

For the first eight data items (publication details, geographical information, research methods used, participant data, variables measured, the discipline of each study), data was extracted from each paper according to the categories and definitions detailed in Appendix A. For each study, a summary phrase was recorded for each finding and conclusion related to L2 aptitude
in that study. Each study was then categorised by keywords from the text in the findings and conclusions and each category related to one of ten themes for L2 aptitude extracted from five theoretical reviews: Dörnyei (2005), Robinson (2012), Sawyer & Ranta (2001), and Skehan (1998, 2002). These ten themes, identified as important across more than one review, are as follows:

1. Purpose and relevance of L2 aptitude
2. L2 aptitude and intelligence
3. L2 aptitude, working memory, and other cognitive processes
4. Measurement of L2 aptitude
5. L2 aptitude and aptitude-treatment interactions (ATIs)
6. L2 aptitude and second-language acquisition processes
7. Structure of L2 aptitude
8. The origins of L2 aptitude and its relationship with the L1
9. L2 aptitude and instruction
10. Stability and untrainability of L2 aptitude

To give an example of how this process worked in practice, a study in the review may have a summary entered in the findings of “MLAT IV correlated with Raven’s Progressive Matrices test”, which in turn would have a summary entered in the conclusions that “language analytic abilities correlates with general cognition”. When searching for studies on L2 aptitude and general intelligence, a keyword search of “general cognition” would return the study given in this example. This study would then be examined and interpreted for its relevance and contribution to research into L2 aptitude and intelligence and reported accordingly in the review. In this way all studies relevant to each theme were identified.
These themes are important to the current study in two ways. First, the themes allow for the
clear organisation of studies and findings to achieve the three objectives of the review (see
section 2.2). Second, each theme is relevant to the current study in one or more ways. For
example, identifying a lack of application of L2 aptitude research (theme 1) in combination
with a clear lack of studies on stability and untrainability (theme 10) motivated the current
study to investigate the efficacy of a training treatment in language analytic abilities.
Language analytic abilities were chosen as a focus, because the results from studies on the
structure of L2 aptitude (theme 7) clearly show the construct is componential, yet most
measurement of L2 aptitude (theme 4) has focussed on aggregate scores as opposed to scores
from sub-tests of individual components or abilities. In this way, the themes were useful in
not only structuring the information of the review, but also for situating, motivating, and
informing the design of the current study.

2.2.2 Study selection

From the initial screening of the keyword database searches, 609 papers were selected and
assessed for eligibility. From these 609 studies, 375 studies were excluded for being
non-empirical (e.g. theoretical reviews, book reviews), book chapters, or not relating to the
prediction of L2 learning. From the remaining 228 papers on the prediction of L2 learning,
another 105 papers were excluded for not utilising an instrument explicitly designed to
measure L2 aptitude or one of its components. The remaining 123 papers were included and
constituted the dataset for the systematic review (see Appendix B for complete list of
references for the review). Figure 2.3 details this process in a flowchart.
2.2.3 Terminology

For the purpose of the review, the term “paper” refers to a publication and the term “study” refers to research where data was collected from a group of participants and used for analysis. A paper can report on more than one study, with each study being recorded separately in the review dataset, as long as the participants and the data differ between the studies. This situation arose where replication or follow-up studies had been conducted. While follow-up studies could have been considered part of a longitudinal data collection and therefore constituted a single study, this categorisation was not used. The follow-up studies recorded almost always had a different research focus and/or different data, e.g. the use of a different
instrument, and did not maintain the complete original sample of participants. In total, the review covers 123 papers and 141 studies.

2.2.4 Results: The general characteristics of L2 aptitude

The aim of the analysis is to summarise the field of L2 aptitude research and identify its boundaries and generalisations. This in turn is to be used to situate, motivate, and inform the design of the current study. Following the guidelines of Pickering and Byrne (2014), the results of the review will be structured as following. Section 2.2.4.1 focuses on authors, publishers, and types of publication most prevalent in L2 aptitude research. Section 2.2.4.2 explores the geographical spread of L2 aptitude research, looking at where the institutions conducting the studies are located and where data are collected from. Section 2.2.4.3 considers the research methods used. Section 2.2.4.4 looks at the participants in L2 aptitude studies and attempts to generalise what the characteristics of the average participant are. Section 2.2.4.5 then turns to the variables in L2 aptitude research including dependent variables, independent variables, as well as the more specific components of L2 aptitude. Following from this, section 2.2.5 examines the findings and conclusions from the 10 themes of L2 aptitude covered by the studies in the review. Each section first summarises the relevant data and then comments upon the results, before contextualising these last within the current study. Where appropriate, further specific analysis are conducted and discussed. (For complete details of the systematic review, see Appendix A.)
2.2.4.1 Publication details of L2 aptitude research

Overall, most L2 aptitude research is published in second language learning journals (see figure 2.4) with an increase in publications since the 1990s (see figure 2.5). Before the 1990s, L2 aptitude publications included reports and a smattering of PhD theses. Since the 1990s, the number of PhD theses has increased, becoming an important source of research for the field.

Figure 2.4 Journals of L2 aptitude research with more than one publication (n > 1)
From the publication data, no long-term research program of L2 aptitude is evident, with very few authors publishing more than one or two studies as first authors (see figure 2.6). The exception is Sparks, Ganschow, and colleagues\(^3\) with approximately 18% of all papers in the review. Sparks is also the author with the most publications of L2 aptitude research overall, with 23 papers including him as an author (not only as first author). This is a potential bias in this review, but also suggests that the Linguistic Coding Deficit Hypothesis (LCDH)\(^4\) of Sparks and Ganschow has the most empirical support. This empirical base for the LCDH contrasts with the relative lack of support for other theories in L2 aptitude (see section 2.1.2 –

---

\(^3\) Co-authors of Sparks and Ganschow included in this review are (in alphabetical order): Artzer, M., Bishop-Marbury, A., Hordubay, D.J., Humbach, N., Javorsky, J., Miller, K., Patton, J., Philips, L., Plageman, M.\(^*\), Pohlmam, J., Siebenhar, D., Skinner, S, Walsh, G.

\(^4\) The LCDH claims that both L1 and L2 acquisition derive from the same general language learning abilities and that problems in a given componential skill of L1 will have negative outcomes for L2 learning in the corresponding skill.
2.1.5). Consequently, the current study is informed by the results from the LCDH but investigates areas of L2 aptitude outside of its direct scope.

![Bar chart showing number of publications by authors](image)

Figure 2.6 First authors of L2 aptitude research with more than one paper (n > 1)

2.2.4.2 Geographical spread

Figure 2.7 is a density map showing the countries in which the institutions of all authors of L2 aptitude papers in the review are located. Countries shaded yellow have a larger number of published papers from authors in that country, while countries shaded red/orange show the opposite.
Figure 2.7 Density map of countries where L2 aptitude research has been conducted

Figure 2.8 is also a density map, this time showing the location of the institutions with which all authors of L2 aptitude papers are affiliated. The colours and the size of the circles both indicate the number of institutions with first authors who have published studies on L2 aptitude. Yellow with a larger circle indicates a larger number of first author publications. Unsurprisingly, the universities of Sparks (College Mt St Joseph), Ganschow, and colleagues (Miami University) in Ohio are the most prominent for L2 aptitude publications.
Finally, figure 2.9 is a density map showing the countries from which data were collected for each L2 aptitude study in the review. From the density maps it is clear that most of what is known about L2 aptitude comes from the (eastern) U.S. context. L2 aptitude is heavily biased towards language learning in the U.S. context with the vast majority of research institutions and data collected being in the United States of America. As a result, the current study seeks to diversify this data set by investigating L2 aptitude issues in the Australian university context, which is comparable to but also distinct from that of North America.
2.2.4.3 Research methods

The research methods of L2 aptitude reflect its psychometric nature. They can be typified as quantitative (95% of studies) not qualitative; longitudinal (68% of studies) more than cross-sectional (31%); observational (67% of studies) more than experimental (33%); interested in relationships between variables (50%) and comparing groups (45%), while also also focussed on variance explained (31%). Statistical analyses most commonly employed (in order) are correlation (45%), ANOVA (28%), multiple regression (26%), and independent $t$-tests (16%).

Since the 1990s, however, research methods have become more sophisticated, with studies exploring more nuanced questions. This can be seen in the increase of cross-sectional and experimental studies, with greater variation in groupings, and an increase in the use of novel
statistical techniques, e.g., structural equation modelling (Cochran, McCallum, & Bell, 2010; Espinosa, 2007; Gardner, Tremblay, & Masgoret, 1997; Gardner & Lysynchuk, 1990; Hwu, Pan, & Sun, 2014; Sasaki, 1993; Sick, 2007; Winke, 2013; Wistner, 2014). This is a promising development for making new findings in L2 aptitude research. The increase in the use of sophisticated research methods and, in particular, novel statistical techniques is needed to unravel the complicated relationship between L2 aptitude and L2 learning. This relationship involves multiple variables interacting with multiple other variables and calls for statistical techniques that can deal with these large numbers of variables with more flexibility than standard linear models, e.g. random forest models (see section 4.6 for details on the statistical analysis used in the present study). However, there is also a noticeable lack of qualitative detailed investigation of individuals in L2 aptitude research, so case studies may be a worthwhile area for future research, e.g. Biedroń & Szczepaniak (2009).

2.2.4.4 Participants

The average participant is 20 years old, male, university educated, and speaks L1 English. This abstraction is drawn from sample sizes appropriate for SLA (median sample size = 90 participants), with gender fairly balanced across studies. Participants across studies are mostly teenagers and young adults from high school and university, with the vast majority speaking L1 English.

In terms of participant background characteristics, the levels of reporting varied. Almost all studies reported on participant numbers (99%), L1s (97%), and education level (81%). The majority of studies reported on participant gender (67%), age range (60%), and average age (53%). However, SES (15%), an important variable in education levels, was rarely reported.
Overall, the data collected from these participants cannot be considered as representative of the population, but rather as representative of a small and select subset of the population, i.e. U.S. high school and university students with English as their L1. The population relevant to L2 aptitude is potentially everyone, so more work is needed to study L2 aptitude with more varied samples.

The lack of consistency in detailed reporting of participant background characteristics makes comparisons across studies challenging. First, it is difficult to compare participant groups as unknown background characteristics may affect the interpretability of results. Second, with such a high concentration of studies on one particular group, any study that wants to explore L2 aptitude issues in more depth needs to be comparable with this group. This results in more studies that focus on high school and university L2 learners. Perhaps when more fundamental questions are answered, more diverse population samples can be investigated. But for now, it may be that for L2 aptitude research to be relevant, it needs to focus on participant groups that are similar to high school and university L2 learner populations, which is why the current study has investigated this group of L2 learners.

**2.2.4.5 Variables in L2 aptitude**

Overall, the variables in L2 aptitude studies tend to be general and aggregate measures, be they dependent variables of overall L2 achievement (figure 2.10), independent variables of general intelligence, or L2 aptitude as a whole. Figure 2.11 shows the most common variables measured⁵ are general cognition (no specific components), attitude (18%), motivation (16%), L1 skills (15%), anxiety (13%), and WM (12%). This suggests that the research is yet to

---

⁵ 18% of studies also included no other variables in their study.
untangle the specifics of L2 aptitude and related constructs, such as general cognition, motivation, and attitudes. A key issue here, of course, is that of measurement. With the best results/correlations in the research coming from aggregate/general measures of L2 aptitude and L2 learning (see section 2.2.5.2.3), the question must be asked: what do the tests really measure? The use of general measures suggests that much remains unknown about what L2 aptitude tests actually measure (Skehan, 1998, 2002).

With few studies focussing, or purporting to focus, on a single component of L2 aptitude, the relationship between specific components of L2 aptitude and other variables in the research, e.g. intelligence and WM, remains unclear. Nevertheless, L2 aptitude research since the 1990s has started to take a more detailed approach to investigating the variables related to L2 aptitude. Coupled with the increase in the number of studies looking at L2 aptitude as a whole as well as its individual components, this suggests that more is being uncovered about the construct of L2 aptitude. With this in mind, the current study is concerned with specific components of L2 aptitude, not just the whole; while many variables (both dependent and independent) identified in this systematic review are not directly investigated in the current study, many of the former (e.g. L2 experience, L1 skills, multilingualism, learning strategies, motivation, and attitude) were used to inform and motivate the latter, specifically, in the design and implementation of the treatment training course. Other variables (e.g. personality, gender) were either too complex and/or did not display a direct relationship with L2 aptitude itself and so were incorporated in the study.
Figure 2.10 Dependent variables in L2 aptitude studies

Figure 2.11 Independent variables in L2 aptitude studies
2.2.5 Results: Findings and limits of L2 aptitude research

Table 2.7 details the ten key themes identified in the theoretical reviews (Dörnyei, 2005; Robinson, 2012; Sawyer & Ranta, 2001; Skehan, 1998, 2002) of L2 aptitude. In the following section each theme will be discussed, both in terms of what was already known from other systematic reviews (e.g. Li 2014, 2015) and what the data from the current systematic review can tell us.

We will first discuss the purpose and relevance of L2 aptitude (section 2.2.5.1), followed by L2 aptitude and intelligence (section 2.2.5.2.1), L2 aptitude and working memory (section 2.2.5.2.2), and the measurement of L2 aptitude (section 2.2.5.2.3). These four themes were not a strong focus of the current review, as it was either impractical, in the case of the purpose and relevance of L2 aptitude, or had already been covered quite comprehensively by Li’s (2015) meta-analysis, in the case of intelligence, working memory, and the measurement of L2 aptitude. In addition, these four themes were not as directly relevant to the current study. Finally, we will discuss L2 aptitude and aptitude-treatment interactions (ATIs) (section 2.2.5.3), L2 aptitude and SLA processes (section 2.2.5.4), the structure of L2 aptitude (section 2.2.5.5), the origins of L2 aptitude and its relationship with the L1 (section 2.2.5.6), L2 aptitude and instruction (section 2.2.5.7), and stability and untrainability (section 2.2.5.8).

Each of these areas will be discussed focusing on what the data from the current review show, as well as highlighting where potential gaps lie. In particular, the issues of the structure of L2 aptitude, its origins and relationship with L1 skills, and stability and untrainability are most relevant to the current study.
Table 2.7 The prevalence of key research themes in L2 aptitude research

<table>
<thead>
<tr>
<th>Theme</th>
<th>Number of Studies</th>
<th>Percentage of all studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>The purpose and relevance of L2 aptitude</td>
<td>141*</td>
<td>*100</td>
</tr>
<tr>
<td>L2 aptitude and intelligence</td>
<td>49</td>
<td>35</td>
</tr>
<tr>
<td>L2 aptitude and working memory</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>Measurement of L2 aptitude</td>
<td>65</td>
<td>46</td>
</tr>
<tr>
<td>L2 aptitude and aptitude-treatment interactions (ATIs)</td>
<td>27</td>
<td>19</td>
</tr>
<tr>
<td>L2 aptitude and second-language acquisition (SLA) processes</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Structure of L2 aptitude</td>
<td>26</td>
<td>18</td>
</tr>
<tr>
<td>The origins of L2 aptitude and its relationship with the L1</td>
<td>29</td>
<td>21</td>
</tr>
<tr>
<td>L2 aptitude and instruction</td>
<td>31</td>
<td>22</td>
</tr>
<tr>
<td>Stability and trainability of L2 aptitude</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

2.2.5.1 Purpose/relevance of L2 aptitude

The term ‘purpose’ in the current study refers to the aim of the L2 aptitude studies reviewed. More specifically, ‘purpose’ has been defined as the explanatory role of L2 aptitude as a variable in a study. This use of purpose contrasts with Dörnyei’s (2005) and Skehan’s (1998) discussions on the purpose of L2 aptitude, e.g. research, selection of students into courses,

* ‘Purpose’ here refers to the aim of the L2 aptitude study, so each study is included in this category, in contrast to Dörnyei’s (2005) and Skehan’s (1998) discussions on the purpose of L2 aptitude (refer to papers for more details).
7 *NB: The review consisted of 141 studies and 123 papers (see section 2.2.3)
individualizing L2 instruction (refer to papers for more details). Consequently, in the current study the most common purpose of L2 aptitude in research has been to explain variance in independent variables in SLA studies and to conduct basic research in exploring the construct of L2 aptitude. There is a clear lack of application of L2 aptitude, despite its potential being identified in the theory (Robinson, 2012; Sawyer & Ranta, 2001). The current study takes up this challenge by investigating a training treatment for a specific component of L2 aptitude (language analytic abilities) that, if successful, could lead to the improvement of language learning skills in participants. This application of L2 aptitude theory could benefit a wide range of users and potentially enable further training treatments for other components of L2 aptitude to be developed.

2.2.5.2 Li’s (2015) meta-analysis

Li’s (2015) meta-analysis of 66 studies and 109 unique samples of L2 aptitude addressed the three themes of: L2 aptitude and intelligence, L2 aptitude and working memory, and the measurement of L2 aptitude. Each theme is discussed below, summarising Li’s (2015) findings and how they compare with the results from the current review.

2.2.5.2.1 Intelligence

Regarding L2 aptitude and intelligence, Li (2015) concluded that while measures of the two constructs overlapped, L2 aptitude was distinct from general cognition. The results from the current review generally concur. From the current review, some studies found that L2 aptitude and intelligence are related (Feenstra, 1967; Sasaki, 1993a, Wesche, Edwards, & Wells, 1982), others that L2 aptitude and intelligence overlap but are separate (Carroll, 1962;
Grigorenko et al., 2000; Sparks, Patton, Ganschow, & Humbach, 2009; Sparks, Humbach, Patton, & Ganschow, 2011), and yet others that have begun to uncover how they are related in L2 learning (Morgan-Short et al. 2015, Sasaki 1993b).

Arguments for L2 aptitude and intelligence being related come from correlational evidence (Feenstra, 1967) and second-order factor analysis, where both constructs are subsumed by a single higher-order factor (Sasaki, 1993a, Wesche et al., 1982). Arguments for the separation of L2 aptitude and intelligence come from the fact that L1 skills account for a greater amount of variance than intelligence measures in correlations with L2 aptitude (Sparks, Patton, Ganschow, & Humbach, 2009) and their separation in a first-order factor analysis (Sparks et al., 2011). Other arguments for the separation of the two constructs come from test validation studies for the MLAT (Carroll, 1962) and the CANAL-FT (Grigorenko et al., 2000), both of which found that the L2 aptitude test scores were better predictors of L2 achievement and that the correlations between the L2 aptitude test scores and intelligence measures were low enough to suggest separation. Studies that have begun to uncover the nature of the relationship between L2 aptitude and intelligence found that L2 aptitude test-takers use the same cognitive processes but employ them differently through different strategies (Sasaki 1993b) and that domain-general abilities were implicated in L2 grammar acquisition (Morgan-Short et al., 2015). These are promising developments in L2 aptitude research in adding much needed detail to our understanding of the construct.

Plainly, general cognition is an important issue in L2 aptitude research, with 49 studies or 35% of all studies in the review including this variable in their research (see figure 2.11). However, in most cases, general cognition was operationalised as either total scores from general intelligence tests (14% of all studies), e.g. Woodcock–Johnson Psycho Educational
Battery (WJPB; Woodcock & Johnson, 1977), or academic aptitude (9% of all studies), e.g. the Scholastic Aptitude Test. So, although the evidence supports some separation of L2 aptitude from general cognition, it seems more research is needed to uncover the nature of the overlap between the two constructs.

In relation to the current study, the evidence for separation of the two constructs suggests that investigating L2 aptitude independently of measures of general cognition is justified. However, the relationship between the two constructs is also in need of further investigation. Findings that show correlations between measures of language analytic abilities and general cognition (Carroll, 1962; Feenstra, 1967; Sasaki, 1993; Wesche et al., 1982) suggest that this is an issue where the overlap between the two constructs is most relevant and, therefore, one of the areas upon which this study will focus (see section 3.2).

2.2.5.2.2 Working memory (WM)

Li (2015) also looked at the relationship between L2 aptitude and working memory (WM). He found that executive working memory is more related to L2 aptitude components than phonological short-term memory (PSTM). The difference between these two types of working memory comes down to processing. Executive working memory involves the storage and processing of information, measured in tests such as Daneman and Carpenter’s (1980) reading span test, while phonological short-term memory only involves storage, measured in tests such as non-word repetition task (Baddeley, 2003). Of the 17 studies of L2 aptitude that included a measure of WM, one included a measure of storage alone and one for processing alone. The other 15 studies included measures of both. This suggests that the role of WM in L2 aptitude may still not be clearly understood, despite Li’s (2015) findings.
Two types of L2 aptitude studies that included WM emerged from the review. The first type of study treated WM as a predictor of L2 achievement independent of L2 aptitude (see Chen, 2013; Erlam, 2005; Fujii, 2005; Robinson, 2005; Serafini, 2013; Tolentino & Tokowicz, 2014; Vandergrift & Baker, 2015; Wistner, 2014; Yilmaz, 2013). Results between these tests varied a lot, with small, medium, and large correlations between certain measures of WM and L2 achievement. The second type of study focussed more on the relationship between the two constructs (see Howes, 1999; Roehr & Gánem-Gutiérrez, 2009; Suzuki & DeKeyser, 2016; Winke, 2013; Yalçin, Çeçen, & Erçetin, 2016). Results from these studies suggest that WM as a measure of memory is distinct from L2 aptitude measures (Roehr & Gánem-Gutiérrez, 2009; Yalçin et al., 2016) but that WM may be implicated in L2 aptitude tests of language analytic ability (Yalçin et al., 2016). This indicates a need to study L2 aptitude independently from WM, a suggestion heeded by the current study. Results from both types of tests suggest that much is still to be learned and understood about how WM and L2 aptitude are related as well as how best to operationalise the two constructs in predicting L2 achievement. For example, where do measures overlap and where are they distinct? What kinds of WM measures should be used with what kinds of L2 achievement measures? With this in mind, we now discuss measurement of L2 aptitude.

2.2.5.2.3 Measurement of L2 aptitude

In addition to intelligence and WM, Li (2015) addressed the issue of measurement of L2 aptitude, finding that overall, or aggregate, scores of L2 aptitude were strong predictors of general proficiency, and that the individual sub-tests measured different components of L2 aptitude. This finding motivated the current study to investigate L2 aptitude both as independent components (sub-test scores) and as a whole (aggregate of sub-test scores). In an attempt to capture the recent questioning of the L2 aptitude construct and its measurement
(Li, 2015; Robinson, 2012; Skehan, 2013; Wen et al., 2017), the current review included categories of predictive validity and construct validity for L2 aptitude studies specifically researching the construct of L2 aptitude itself.

Fifty-eight, or 41%, of all studies had findings that related to the predictive validity of L2 aptitude. From these 58 studies, four key findings emerged. First, a lot of support was found for L2 aptitude measures as a valid predictor for L2 achievement, both for specific tests (Bell & McCallum, 2012; Carroll, 1962; Dahlen & Caldwell-Harris, 2013; Ehrman, 1994; Hanna, nd; and Phillips, 1998 for MLAT; Pimsleur, Sundland, & McIntyre, 1963 for PLAB; Asher, 1972 for HABLA & AFLAAT/DLAB; Sasaki, 1993b for LABJ; Grigorenko et al., 2000 for CANAL-FT; and Artieda & Muñoz, 2016 for LLAMA D) and for L2 aptitude in general (Adamu, 1990; Ahn, 2012; Carroll, 1962; Carroll, 1967; Cochran et al., 2010; Ehrman & Oxford, 1995; Espinosa, 2007; Feenstra, 1967; Granena, 2014; Hughes, 2010; Sasaki, 1993a, Theivananthampillai & Baba, 1984; Wangsotorn, 1987).

Second, some studies highlighted the limitations of the predictive validity of L2 aptitude measures. Winke (2013) found that the MLAT was not a good predictor of advanced L2 achievement, while Ehrman (1994) found that the MLAT was better at discriminating between learners with low L2 aptitude test scores than between learners with high L2 aptitude test scores. Gardner and Lysynchuk (1990) found that L2 aptitude predicts short term achievement, while Granena (2014) found that L2 aptitude tests were measuring L2 aptitude for explicit more than implicit learning.

Third, various issues in operationalizing L2 aptitude were also raised in multiple studies. Brecht, Davidson, and Ginsberg (1993); Carroll (1962); and Curtin, Avner, and Smith (1983) found that the validity of both the MLAT and PLAB scores was sensitive to the
criterion/dependent variable. Howes (1999) also found that testing needed to take place before L2 learning for MLAT scores to hold their validity, while Davies (1971) found that the type of instruction for L2 learning was also an important issue for the predictive validity of L2 aptitude measures. Sparks and Patton (1999); Sparks, Patton, Ganschow, and Humbach (2009); Sparks, Philips, and Javorsky (2002); Sparks, Philips, Ganschow, and Javorsky (1999) also found that L2 aptitude was insufficient to predict L2 achievement without including measures of L1 skills.

Finally, several studies also found other significant predictors of L2 achievement in conjunction with L2 aptitude. Horwitz (1980) found that conceptual level, defined as “an information processing variable which measures an individual's cognitive complexity and interpersonal maturity” (p.2), was a key predictor of L2 achievement above L2 aptitude. However, conceptual level differentially predicted L2 communicative competence, while the MLAT IV accounted for linguistic competence. Damiano (1987) found that Hill’s educational model of cognitive styles (Hill 1981 as cited in Damiano 1987:9) was also an important predictor of overall L2 achievement. Morgan-Short, Faretta-Stutenberg, Brill-Schuetz, Carpenter, and Wong (2014) and Morgan-Short, Dneg, Brill-Schuetz, Faretta-Stutenberg, Wong, and Wong (2015) found that declarative memory was an important predictor of early L2 grammar acquisition, whereas procedural memory was an important predictor of late L2 grammar acquisition. Equally, and unsurprisingly, Sparks and Ganschow’s research team found that not only were L1 skills an important predictor of overall L2 achievement, but also of L2 aptitude itself (Sparks, Ganschow, & Patton, 1995; Sparks, Ganschow, Patton, Artzer, Siebenhar, & Plageman, 1997; Sparks et al., 2011). Chen (2001) and Sparks and Ganschow
(1995) also found that L2 difficulties, the negative spectrum of L2 aptitude, can also be predicted.

So what does this tell us about L2 aptitude measurement? First, L2 aptitude measures are valid, but their operationalization may be more restricted than it seems at first. In particular, criterion/dependent variables need to be carefully considered and matched with the L2 aptitude measure of choice. Second, the range of other cognitive variables implicated in L2 learning, e.g. declarative and procedural memory or L1 skills, raises questions about the construct validity of L2 aptitude. Are the more traditional constructs of L2 aptitude based on the MLAT too restrictive (see section 2.1.2 for discussion)? For example, the CANAL-FT has a completely different theoretical construct to the MLAT and yet it still proved to be a valid test of L2 aptitude. Developments like this are what have led to the more recent efforts to reformulate the construct of L2 aptitude (e.g. Skehan, 2002; Robinson, 2012; Wen et al., 2017). With this in mind, we now turn to the data from the review regarding the construct validity of L2 aptitude.

Seven, or 5%, of all L2 aptitude studies focussed on the construct validity of L2 aptitude. Many of these studies have already been discussed. Grigorenko et al. (2000), Sasaki (1993), Wesche et al. (1982) all challenge the specificity of L2 aptitude by linking it to general cognition, and Winke (2013) found that the MLAT was not predictive of L2 ultimate attainment. Other studies have challenged the construct of L2 aptitude in different ways. VanPatten & Smith (2014) looked at L2 aptitude from a processing perspective and argued that grammatical sensitivity was not central to SLA due to the lack of correlations in their study. However, it should be noted that their criterion/dependent variable was based on L2 processing, which perhaps does not match very well with the MLAT IV. Garcia (1996) found
support for the construct of aptitude in Gardner’s (1985) socio-cultural model, while G. Robinson (1975) found that L2 learning experience was more influenced by social variables than L2 aptitude, challenging the importance of L2 aptitude as a predictor of L2 achievement.

Two other studies also had findings related to the construct validity of L2 aptitude. Carroll (1962) identified that only a limited number of abilities was needed to validly and reliably predict L2 achievement, restricting the range of the L2 aptitude construct to four components, while Roehr (2007) found that the language analytic ability component of L2 aptitude may in fact have to be expanded to include the use of metalinguistic knowledge to identify, label, link, and match L2 grammatical categories and their relationships to other L2 grammatical categories. Studies such as Roehr’s (2007) suggest that language analytic ability may involve more than just grammatical sensitivity and inductive language learning ability as measured by the MLAT IV (Carroll & Sapon, 1959) and the PLAB 4 (Pimsleur, 1966), respectively (see section 3.2 for discussion). This reflects Robinson’s (2012) argument that a minimally sufficient approach to L2 aptitude testing is not necessarily useful in understanding how L2 aptitude abilities relate to SLA processes and that, consequently, we need new tests of L2 aptitude abilities that measure the construct at finer levels of detail.

Despite these challenges to the construct validity of L2 aptitude, the predictive validity of current L2 aptitude tests seems to be strong enough to support the construct, giving credence to the assertion that L2 aptitude is what L2 aptitude tests measure (Dörnyei, 2005). However, at a more detailed level, much is still to be learned about the individual components of L2 aptitude, which is why the current study includes measures of both sub-tests and overall L2 aptitude test performance in its analysis. The issue of construct validity also relates to the
themes of L2 aptitude and aptitude-treatment interactions, L2 aptitude and SLA processes, and the structure of L2 aptitude, to which we now turn.

2.2.5.3 Aptitude-treatment interactions

Snow (1991:205) defines aptitude-treatment interactions (ATIs) as the situation in which “individuals differ in their readiness to profit from a particular treatment at a particular time”. Using this definition, the current review found that 27, or 19%, of all studies had evidence for ATIs in L2 aptitude, although it is important to note that not all these studies were designed as ATI studies in the technical sense (see Snow, 1991 for explanation of ATI research designs). Support for ATIs in L2 aptitude came from two types of studies.

Studies that compared groups of learners with differing L2 aptitude profiles and/or L2 learning treatments found that different L2 aptitude components differentially predicted L2 achievement for a given treatment (Artieda & Muñoz, 2016; Bain, McCallum, Bell, Cochran, & Sawyer, 2010; Chen, 2013; Davies, 1971; Fujii, 2005; Hwu, Pan, & Sun, 2013; Jiang & Xiao, 2014; Robinson, 2005; Sachs, 2010; Serafini, 2013; Stefanou & Révész, 2015; Suzuki & DeKeyser, 2016; Takeuchi-Furuya, 1993; Tolentino & Tokowicz, 2014; VanPatten & Borst, 2012; Yalçın & Spada, 2016). For example, Hwu, Pan, and Sun (2013) found that low-aptitude learners with deductive instruction outperformed those with explicit-inductive instruction, while for all learners there was no difference between the treatment groups. These results are suggestive of ATIs in L2 learning.

Studies that examined the variables of success for a given L2 learning situation found that differing profiles of aptitude, often with other variables, predicted success for L2 achievement (Ahn, 2012; Allen & Natelson, 1977; Brecht et al., 1993; Damiano, 1987;
Demuth & Smith, 1987; Garcia, 1996; O'Maley, 1993; Wangsotorn, 1987; Vandergrift & Baker, 2015). Ahn’s (2012) study illustrates this point, finding that higher grammatical sensitivity led to greater levels of noticing, which enabled learners to benefit more from recasts. These results are also suggestive of ATIs in L2 learning.

If we read these studies as investigations of success in a given treatment, then the results can be considered to support ATIs and, by extension, Robinson’s (2001) Aptitude Complex Hypothesis. However, two studies contradicted this trend. Erlam (2005) found that deductive instruction may have neutralized the effect of L2 aptitude on L2 learning, while Dahlen & Caldwell-Harris (2013) found no differences across three different treatments for L2 aptitude levels. While these results certainly pose interesting questions about the relationship between L2 aptitude and instruction, they do not invalidate the other findings of ATIs in L2 learning. However, they do highlight the need for more detailed research into the interactions between L2 aptitude, instruction, and L2 learning. The increasing use of ATI studies holds much promise for L2 aptitude research. All three of the factors previously mentioned informed the design of the current study, despite its not being an actual ATI study. More importantly, the current study was influenced by ATI theory in (a) considering that L2 aptitude was not stable and was, therefore, trainable (see section 3.1.3); (b) presuming that different types of learners would benefit from instruction differentially; and (c) treating the components of L2 aptitude as important factors in L2 learning outcomes beyond their contribution to aggregate test scores.
2.2.5.4 **L2 aptitude in SLA processing stages**

From a certain perspective, Skehan’s (2002) model of L2 aptitude and stages of SLA is compatible with an ATI view of L2 aptitude (Robinson, 2012). Each stage of L2 acquisition can be viewed as presenting the learner with a different task that draws upon different components of L2 aptitude for successful learning. Consequently, there is a crossover between the studies above in support of ATIs and those supporting Skehan’s (2002) model of L2 aptitude and SLA. In total, six, or 4%, of studies found evidence in support of Skehan’s (2002) model.

Studies that compared L2 learners of differing levels found that different components of L2 aptitude were more predictive of L2 achievement at different stages of L2 proficiency (Artieda & Muñoz, 2016; Serafini, 2013; Yalçın & Spada, 2016). For example, Yalçın and Spada (2016) found that phonological abilities were more important at earlier stages of L2 learning, while language analytic abilities were important at all stages of L2 learning.

Studies that focussed on learners of a specific L2 level found that specific components of L2 aptitude were more important for that particular level (Demuth & Smith, 1987; Fujii, 2005; Wistner, 2014). For example, Fujii (2005) found that intermediate to high intermediate L2 learners of English with lower levels of language analytic abilities and higher orientation to form showed greater lexical variety. Both sets of studies support Skehan’s model in general as well as some of its specific predictions, e.g. that phonological abilities are more implicated in the initial stages of L2 learning. Conversely, this also predicts that language analytic abilities are less implicated in the initial stages of learning, such that L2 learners are practicing phonological skills more than language analysis skills. Therefore, if any training effects were present in L2 aptitude scores as a result of L2 learning, Skehan’s (2002) model would predict
that phonological abilities would show greater increases in L2 aptitude test scores than language analytic abilities. The current study followed this reasoning and investigated the effects of a training course on language analytic abilities for L2 learners beginning Spanish, as this component of L2 aptitude was the least implicated in the initial stages of L2 learning.

However, some contradictions across studies also emerge. The finding from Yalçın and Spada (2016) regarding the importance of language analytic abilities at all levels of L2 learning does not match those of Fujii (2005) and Wistner (2014), for whom language analytic abilities were not important predictors of L2 aptitude for intermediate to high intermediate L2 learners of English. This is an interesting situation, especially given the strength of language analytic abilities in predicting overall L2 achievement (Li, 2015). Two factors may be at play here. The first relates to the differences in measures: Yalçın and Spada (2016) used the LLAMA test (Meara, 2005), Fujii (2005) the LABJ (Sasaki, 1991), and Wistner (2014) the Lunic Language Marathon (Sick, 2007). Equivalencies across these tests for their predictive power and the abilities they measure do not exist, so it is impossible to currently resolve this issue. The second factor relates to the context of learning. Yalçın and Spada (2016) looked at adults in Spain at a language school, while Fujii (2005) and Wistner (2014) both examined L1 Japanese learners of L2 English in a Japanese university context. Instruction, possibly also implicating culture, is another variable here unaccounted for in comparing the results of these studies. Despite these inconsistencies in findings, there is nevertheless some compelling evidence for Skehan’s (2002) model of L2 aptitude and SLA, but more research is required.

### 2.2.5.5 Structure of L2 aptitude

From the review, 26, or 18%, of studies investigated or had findings relating to the structure of L2 aptitude. A range of statistical methods was used to investigate the structure of L2
aptitude, including multiple regression (see Damiano, 1987; Detschelt, 1993; Espinosa, 2007; Gilleece, 2006; Horwitz, 1980; Howes, 1990; Morgan-Short et al., 2014; Sparks, Ganschow, & Patton, 1995; Sparks, Ganschow, Patton, Artzer, Siebenhar, & Plageman, 1997; Sparks et al., 2011), principal components analysis (see Feenstra, 1967; Sparks, Ganschow, Patton, Artzer, Siebenhar, & Plageman, 1997; Sparks et al., 2011; Roehr, 2007; Yalçin et al., 2016), and structural equation modelling (see Espinosa, 2007; Sasaki, 1993). These studies can be grouped into two main types of studies: those that investigate the internal structure of L2 aptitude and those that investigate where L2 aptitude ends and other cognitive abilities begin.

Sawyer and Ranta (2001) explicitly ask the question of whether L2 aptitude is unitary or componential. From the very beginning, Carroll (1962) viewed L2 aptitude as componential and theorised that it consisted of four components: phonetic coding, associative memory, grammatical sensitivity, and inductive language learning (see section 2.1.2 for details). However, in practice, L2 aptitude tests are mostly combined into a single general score of L2 aptitude and used to predict a single score of general L2 proficiency. Thus, this question of structure is an important issue in L2 aptitude studies. From the review, the evidence clearly supports the componential view of L2 aptitude (Artieda & Muñoz, 2016; Sasaki, 1993a, Sparks et al., 2011), which is also the view adopted for the current study and the reason why both sub-test and aggregate L2 aptitude test scores were treated as meaningful variables.

However, many studies also looked at where L2 aptitude stops and other cognitive abilities start. The relationship between L2 aptitude and intelligence is complex, generally supporting the view that the two constructs are distinct but overlap (Damiano, 1987; Feenstra, 1967; Sasaki, 1993a, Sparks et al., 2002; Sparks et al., 2011) (see section 2.2.5.2.1 for more details). Working memory (WM) has also been discussed previously (see section 2.2.5.2.2), with the
general view being that WM is a separate construct from L2 aptitude that is nevertheless measured to some degree in certain tests of L2 aptitude (Fujii, 2005; Howes, 1999; Yalçin et al., 2016). For other constructs of cognitive abilities, a similar pattern of results was found, in which separation from and some level of overlap with L2 aptitude recurred. This separation and overlap was found for musical aptitude (Gilleece, 2006), cognitive styles (Damiano, 1987), L1 skills (Sparks, Ganschow, & Patton, 1995; Sparks, Ganschow, Patton, Artzer, Siebenhar, & Plageman, 1997; Sparks et al., 2002; Sparks et al., 2011), and metalinguistic knowledge (Alderson et al., 1997; Roehr, 2007; Wistner, 2014). However, Alderson, Clapham, and Steel (1997) and Wistner (2014) found evidence to indicate a separation between L2 aptitude and metalinguistic knowledge, despite suggestions of an indirect theoretical link (see section 3.2.2 for discussion). In yet other areas of cognition, Horwitz (1980) found that while grammatical sensitivity was the best predictor of L2 knowledge, conceptual level was the best predictor of L2 communicative ability. Sparks and Ganschow’s research team found links between L1 skills and L2 aptitude (Sparks, Ganschow, & Patton, 1995; Sparks, Ganschow, Patton, Artzer, Siebenhar, & Plageman, 1997, Sparks et al., 2002; Sparks et al. 2011; see section 2.2.5.6 for more details). Finally, Artieda & Muñoz (2016) and Morgan-Short, Faretta-Stutenberg, Brill-Schuetz, Carpenter, and Wong (2014) both found explicit and implicit learning dimensions for L2 aptitude. These studies paint a complex picture of L2 aptitude in relation to other constructs of cognitive abilities and suggest that the structure of L2 aptitude may be entwined with other areas of cognition.

Other studies pose questions yet to be answered. Detschelt (1993) concludes that the relationship between L2 aptitude and learning modalities needs clarifying; Sheffield (1993) suggests that L2 aptitude may contain a heritable component; Yalçin, Çeçen, and Erçetin (2016) claim that WM is, in fact, separate from L2 aptitude and also question Granena’s
L2 aptitude structure of implicit and explicit components; and Cummins and Gulutsan (1975) indicate that L2 aptitude may in some part be related to the ability of quick set extinction – the ability to switch between (autonomous) linguistic systems (Cummins & Gulutsan, 1975) – requiring a more complex cognitive structure than current models take into account.

In short, the research supports the view of L2 aptitude as an umbrella term for a range of cognitive abilities that underlie L2 learning. Currently, the L2 aptitude test paradigm groups these into three components: phonological abilities, memory (Skehan, 1998, 2002), and language analytic abilities (combining Carroll’s grammatical sensitivity and inductive language learning, see section 3.1 and Skehan, 1998 for details). The lack of definitive answers and the novel studies in this area suggest that there is much work required to identify exactly what L2 aptitude is and to untangle it from related constructs such as intelligence and working memory.

2.2.5.6 Origins and relationship with L1

Discussing the origins of L2 aptitude, Carroll (1973) hypothesized that it was probably genetic and a remnant of individual differences in L1 acquisition. However, few studies seem to have investigated this area of L2 aptitude research with only eight, or 6%, of studies displaying findings related to these topics. Only one study investigated the heritability of L2 aptitude. In a twin study of adult foreign language learning, Sheffield (1993) found that there is some genetic influence in articulatory ability, which suggested a possible heritable component to L2 aptitude. This is a vastly under-examined area and warrants further research.
The remaining seven studies related to the connection between L1 and L2 aptitude. Sparks and Ganschow’s research team were the most active researchers in this area, due to its relation to the Linguistic Coding Deficit Hypothesis (LCDH). Importantly, they found that L1 skills are an important predictor of L2 aptitude (Sparks, Ganschow, & Patton, 1995; Sparks et al., 2011), supporting Carroll’s (1973) hypothesis that L2 aptitude is a remnant of L1 acquisition. Further support comes from other studies showing a relationship between L1 skills and L2 aptitude (Feenstra, 1967; Politzer & Weiss, 1969; Vandergrift & Baker, 2015).

As previously mentioned, Sparks and Ganschow’s LCDH is premised on the link between L1 skills and L2 aptitude (Sparks, 1995). In total, 21, or 15%, of all studies in the review had findings that supported the LCDH (Bain et al., 2010; Chen, 2001; Downey, Snyder, & Hill, 2000; Ganschow et al., 1991; Phillips, 1998; Sparks & Ganschow, 1995, 2011; Sparks & Patton, 1999; Sparks, Ganschow, Javorsky, Pohlman, & Patton, 1992a; Sparks, Ganschow, Pohlman, Skinner, & Artzer, 1992; Sparks, Ganschow, Patton, Artzer, Siebenhar, & Plageman, 1997; Sparks et al., 1998; Sparks et al., 1999; Sparks et al., 2008; Sparks, Patton, Ganschow, & Humbach, 2012). With such a wealth of empirical support, it seems more than plausible that L2 aptitude is related to L1 skills. This clearly indicates that if testing participants with a variety of L1 backgrounds, then the L2 aptitude must not be L1 specific, e.g. the MLAT (Carroll & Sapon, 1959) or the PLAB (Pimsleur, 1966). This is why in the current study a language-neutral test of L2 aptitude, the LLAMA tests (Meara, 2005), was employed. However, more work is required to explain this link and what it means for L2 aptitude. One area in which some research has already been done, using the LCDH as a theoretical base, is in L2 instruction for at-risk L2 learners (see 3.1.2.1 for details), discussed further in the following two sections.
2.2.5.7 Instruction

As part of his theoretical musings on L2 aptitude and L2 learning, Carroll (1962) proposed the model of school learning. Central to the model is the role of instruction. In fact, Carroll (1981) also claims that L2 aptitude, as operationalised primarily by the MLAT, relates to formal, or instructed, L2 learning. In total, 31, or 22%, of all studies in the review had findings that related to the issue of instruction and L2 aptitude. These studies fell into three main groups: those whose findings showed implications for L2 aptitude on instruction, those whose findings showed implications for instruction on L2 aptitude, and those whose findings raised issues about the relationship between L2 aptitude and instruction.

A wide range of findings emerged with implications for L2 aptitude on instruction. Much of the research uncovered specifics of how L2 aptitude operates in L2 learning (see Ahn, 2012; Downey et al., 2000; Gajar, 1987; Granena & Long, 2012; Roehr, 2007; Stefanou & Révész, 2015; Sparks & Patton, 1999; Sparks, Ganschow, Javorsky, Pohlman, & Patton, 1992a; Sparks et al., 2008; Roehr & Gánem-Gutiérrez, 2009; Dahlen & Caldwell-Harris, 2013). For example, regarding L2 grammar learning, Ahn (2012) found that grammatical sensitivity predicted learning from recasts as a result of increased noticing, while Roehr (2007) found data to suggest that language analytic abilities may involve processes of categorizing grammatical units and learning the relations between such categories. Other studies found that approach to learning was also connected to L2 aptitude (see Suzuki & DeKeyser, 2016; Politzer & Weiss, 1969; Bartley, 1968). For example, Suzuki and DeKeyser (2016) found that language analytic abilities were more important for distributed learning than massed learning. Explicitness of instruction and L2 aptitude also emerged as an important area (see Cox, 2013; VanPatten, Collopy, Price, Borst, & Qualin, 2013; Wistner, 2014; Tolentino & Tokowicz, 2014). For example, Cox (2013) and Wistner (2014) found that language analytic abilities are
most beneficial for explicit instruction. Related to this, Artieda and Muñoz (2016), Granena (2013), and Robinson (1994) all found evidence of L2 aptitude for explicit and implicit learning, which obviously has important consequences for the role of instruction in L2 learning. Regarding the future of L2 instruction, three studies showed evidence to support the potential of individualized learning (Allen & Natelson, 1977; Hwu et al., 2014; Robinson, G. 1975).

However, the relationship between L2 aptitude and instruction is not one way. Many studies also found that instruction has implications for L2 aptitude. Instruction may actually mediate the effects of L2 aptitude for L2 learning (Erlam, 2005; Jiang & Xiao, 2014) or may even lead to increased L2 aptitude levels (Sparks & Ganschow, 1993; Sparks, Ganschow, Artzer, & Patton, 1997). However, instruction may also have an impact on the measurement of L2 aptitude (Davies, 1971; Howes, 1999), which may also be confounded by other variables such as learning modality (Detschelt, 1993) and motivation (Sick, 2007). These findings suggest that L2 aptitude is influenced by (explicit) instruction. This, in turn, suggests that L2 aptitude may be trainable to an extent, the proposition under investigation in the current study (see section 3.3).

Once again, a very complex picture of L2 aptitude emerges, this time in regards to instruction. L2 aptitude can impact on the efficacy of instruction and, in turn, instruction may have effects on L2 aptitude. This scenario is exactly what to expect with ATIs and highlights the need for specificity when reporting research details regarding the L2 learning of participants. Unfortunately, this is an area in L2 aptitude research that is under-examined. Of 132 studies involving some sort of instructed learning, sixty-five, or 49%, of these studies did not report on the instructional method of L2 learning participants had undertaken and outside
of laboratory studies (12% of all studies), specifics regarding instruction were also lacking. For the relationship between L2 aptitude and instruction, including any ATIs, to be uncovered, more detail in the reporting of the L2 learning of participants is required.

2.2.5.8 Stability and untrainability

Skehan (1998) explains that two of the key assumptions of L2 aptitude are stability and untrainability. These terms are related in that the latter is dependent upon the previous – if L2 aptitude is stable, then it follows that it is also untrainable. However, the two primary focuses has been on the stability of L2 aptitude (Carroll, 1973, 1981; Rogers et al., 2017; Singleton, 2017; Skehan, 1998, 2002) and not on untrainability. Plainly, the presumption is that until L2 aptitude can be considered unstable, then it cannot be considered trainable. However, if L2 aptitude is not stable, then it does not necessarily follow that it is also trainable. Consequently, stability and untrainability are related but not equivalent terms. However, for the purposes of this thesis, I will continue to use these two binaries more or less indistinctly, consistent with the rest of the field.

Skehan’s discussion on the assumptions of stability and untrainability follows on from Carroll (1981), who claims he had seen no evidence to support that L2 aptitude can be modified in any meaningful way over the long-term. But as Snow (1991:206) states “modifiability and continuity of aptitude differences represent questions for research, not assumptions”. Unfortunately, not many studies have focussed on these questions, with only eight, or 4%, of all studies reporting on stability and only two studies reporting on the trainability of L2 aptitude. The current study attempts to fill this gap by investigating both the stability and trainability of L2 aptitude.
Of the eight studies with findings relevant to the stability of L2 aptitude, half were cross-sectional studies and half were longitudinal. Regarding stability, the four cross-sectional studies found that participants who had previous language learning experience before taking L2 aptitude tests outscored those who had not (Allen & Natelson, 1977; Planchon & Ellis, 2014; Thomson, 2009, 2013). The four longitudinal studies found that L2 learners’ scores on the MLAT increased after completing one year’s instruction (Demuth & Smith, 1987; Ganschow & Sparks, 1995; Sparks, Ganschow, Pohlman, Skinner, & Artzer, 1992) and also after two years’ instruction (Sparks, Ganschow, Artzer, & Patton, 1997). Contrary to the assumption of stability, these studies suggest that L2 aptitude is dynamic and improves with language learning experience, both over the short- and long-term.

Regarding trainability, only two studies explicitly investigated this area. Politzer and Weiss’s (1969) found that L2 aptitude test scores did increase above the average test-retest scores, but their results failed to demonstrate that explicit training in L2 aptitude abilities was more effective than the experience of L2 learning alone. Gain scores were also not statistically significant. Yeni-Komshian (1965) found that training in phonological abilities led to an increase in aptitude test scores, but primarily for those who were already of a ‘high’ level, i.e. approximately 80% and above, in the pre-test.

Evidently, Carroll’s assumption of stability has not been thoroughly tested in research, with only eight studies on stability, and two for trainability. Results from two training studies are not enough to determine that L2 aptitude is untrainable. Negative results may have two possible causes: (1) L2 aptitude is stable and untrainable, or (2) the training implemented was ineffective. Plainly, further research is required; it is precisely in this area that our current study has the most to contribute to the field of L2 aptitude (see chapter 3).
2.2.6 Conclusions for the systematic review of L2 aptitude research

In summary, based on the results of our systematic review of research into L2 aptitude, what generalizations can be made, what limitations can be identified, and what areas emerge as requiring further research?

In terms of generalizations, the research shows that L2 aptitude is componential; overlaps with and is also distinct from general cognition, working memory, and other constructs; is a valid predictor of L2 learning when used with appropriate criterion measures; shows evidence of aptitude-treatment interactions and of being differentially implicated in different stages of SLA; is related in some way to L1 abilities; interacts with instruction in L2 learning; and shows evidence of being sensitive to experience.

There are some strong limitations to these generalizations. First, these results are derived mainly from high school and university L2 learners from the United States of America who speak English as their first language. Second, most research has been conducted using aggregate measures of L2 aptitude, often with aggregate measures of other constructs, e.g. general intelligence. Only recently have studies started to use more specific measures of L2 aptitude and related constructs. However, more detailed understandings of L2 aptitude are still limited. Third, research designs have been almost exclusively quantitative. Consequently, details of individual experiences, a potentially valuable source of data, are lacking. However, the review also shows some promising changes in trends in current L2 aptitude research which will hopefully address some of these limitations.

The results of this systematic review also suggest that in many ways, L2 aptitude is still not very well understood and that many of its aspects remain unclear. What is clear, however, is that three issues above all others remain most under-researched. First, the relationship
between L1 skills and L2 aptitude remains relatively under-examined, although there has been a significant amount of relevant L2 aptitude research carried out by Sparks, Ganschow, and team in investigating the LCDH. Second, while the origins of L2 aptitude are almost completely unresearched, this issue poses substantial challenges to empirical research. Third, and most importantly for the purposes of our current study, the related issues of stability and untrainability in L2 aptitude remain under-researched, particularly untrainability. Given that these are fundamental assumptions of much L2 aptitude measurement (Skehan, 1998, 2002) and therefore research, more research is needed to clarify the existing data that currently challenge this traditional view of L2 aptitude. The current study aims to contribute to filling this gap in the research, although given the limitations of time and resources, only the issues of stability and untrainability will be addressed.

2.3 What is L2 aptitude: summaries and conclusions

In an attempt to answer the question ‘what is L2 aptitude?’ this chapter has considered both theoretical and empirical perspectives. Theoretically, L2 aptitude is based firmly in the reality of L2 aptitude tests. Much of this theory is derived from the MLAT, its development, and Carroll’s post-hoc theorising. However, this theoretical perspective is limited by issues with work-sample tests. As a result, more recent models of L2 aptitude have been proposed, none of which have as yet been completely validated. Nevertheless, cognitive perspectives that adopt an aptitude-treatment interaction approach seem to offer a more holistic view of L2 aptitude with more explanatory power than Carroll’s more restricted, traditional view.

The results of a systematic review suggest that from an empirical perspective, in many ways L2 aptitude is still not very well understood. Usefully, perhaps, this review suggests that L2
aptitude sample sizes tend to be reasonable as well as being balanced for gender; measures of L2 aptitude appear to be valid in relation to appropriate dependent variables; and since the 1990s studies have become more frequent, targeted, and more sophisticated. However, problems uncovered in this review include bias in research institutions, data collection, and types of participants; inconsistent reporting on background information of participants and the types of instruction undertaken; a lack of detail in explaining the structure of L2 aptitude and its relationships with other constructs; and a lack of detailed explanations for many fundamental questions.

One issue that arises in both theoretical and empirical perspectives is the assumption of stability and untrainability. This question in L2 aptitude has been grossly neglected in the research (Robinson, 2012; Skehan, 2002; Sternberg, 2002), especially in the light of theoretical developments since the 1960s. With new ways of understanding L2 aptitude more holistically (e.g., Sternberg, 2002) and some researchers questioning Carroll’s original thinking (Sáfár & Kormos, 2008; Singleton, 2014, 2017; Wen et al., 2017), now seems an appropriate time to revisit the issues of stability and untrainability in L2 aptitude.
CHAPTER 3: INVESTIGATING THE STABILITY OF SECOND LANGUAGE APTITUDE AND THE TRAINABILITY OF LANGUAGE ANALYTIC ABILITIES

The previous chapter identified stability in L2 aptitude and its correlate, untrainability, as a gap in the research. The current chapter, which is divided into two main sections, explores this issue in more detail. Section 3.1 discusses the theory and research relating to stability and untrainability of L2 aptitude. Section 3.1.1 critically reviews Carroll’s theoretical assumptions in this regard. Section 3.1.2 then examines the empirical evidence of stability and untrainability in L2 aptitude. Section 3.1.3 reviews alternative explanations of stability and untrainability from generalist cognitive theories of L2 aptitude. Section 3.1.4 summarises these perspectives and identifies the key factors relevant to the current study. The second part of this chapter (section 3.2) explores the issue of the trainability of language analytic abilities, an understudied component of L2 aptitude (Skehan, 2002). In particular, this construct of language learning abilities is investigated with the goal of developing a training course informed by theory to investigate its trainability. Section 3.2.1 discusses the terminology and problems with associated with defining language analytic abilities. Section 3.2.2 reviews what is known about language analytic abilities and discusses both tests of language analytic abilities (section 3.2.2.1) and findings from relevant studies (section 3.2.2.2). Section 3.2.2.3 summarises and identifies the need for a more detailed explanation of language analytic abilities for the purposes of training. Section 3.2.3 then outlines a novel explanation of language analytic abilities with an overview of how they might operate in

---

8 Skehan (2002) uses the term ‘language analytic ability’, in the singular, to refer to this component of L2 aptitude. However, because the current thesis views this component as an umbrella term for a range of individual abilities (see section 3.2.3) the term ‘language analytic abilities’, in the plural, will be used in this thesis.
second language acquisition (section 3.2.4). An explanation is also provided of how this novel view of language analytic abilities is used in the current study to develop a training course. Section 3.2.5 summarizes the discussion on language analytic abilities and the chapter concludes by articulating the research questions for the current study (section 3.3).

### 3.1 Stability and untrainability of second language aptitude

The assumption that second language aptitude is stable, and therefore untrainable\(^9\), is based on Carroll’s theories (Carroll, 1973, 1981) as well as evidence that shows a link between L2 aptitude and L1 development (Skehan, 1986) and L1 skills (Feenstra, 1967; Sparks, Ganschow, & Patton, 1995; Sparks et. al., 2011; Vandergrift & Baker, 2015). However, this assumption is relatively under-researched (see section 2.2.5.8) and has come under closer scrutiny and criticism in recent times (Robinson, 2012; Rogers et al., 2016; Rogers et al., 2017; Sáfar & Kormos, 2008; Singleton, 2014, 2017). In particular, it is important to note the usage of these two terms.

As the focus of this thesis, the current chapter begins to explore the issues of stability and untrainability of L2 aptitude by critically reviewing the basis of Carroll’s theoretical assumptions of stability (section 3.1.1), thereby extending the discussion from the previous chapter.

\(^9\) Again, it is important to note that the literature in general does not deal with these two issues separately. Put otherwise, L2 aptitude must be construed as not stable before the question of if it is trainable or not can be investigated. Consequently, for the purposes of this thesis, I will continue to use these two binaries more or less indistinctly, consistent with the rest of the field (see 2.2.5.8).
3.1.1 Carroll’s theoretical assumptions

The assumption of stability of L2 aptitude is based on Carroll’s conceptualisation of the construct (Skehan 1998, 2002, see section 2.1.2). Carroll (1973) speculates that L2 aptitude is a function of the remnants of innate L1 acquisitional abilities that decline differentially in individuals for an unknown reason. This implies that L2 aptitude is also innate. Carroll (1981) then concludes that L2 aptitude must have a fundamental biological basis, which is only indirectly evident in learning. In other words, Carroll seems to regard L2 aptitude as a composite of first-order abilities and also seems to imply that these abilities are limited by the capacity of the brain. The resulting assumption, then, is that L2 aptitude is relatively fixed, stable over the long term, and resistant to significant improvement (Carroll, 1981). Interestingly, if we follow Carroll’s reasoning and assume that L2 aptitude is the result of fading L1 acquisitional abilities, then we might also assume that L2 aptitude could deteriorate over time.

If L2 aptitude is stable, then it logically follows that it is also untrainable. While Carroll (1981) concedes that short-term increases in L2 aptitude scores are possible, such gains should return to a long-term baseline over time. Carroll (1962) also points out that trainability of L2 aptitude would need to prove not only that L2 aptitude scores can increase but that this increase would translate into increased L2 learning. To some extent, Carroll’s views here appear to contradict the initial assumption that L2 aptitude is set in stone. To date, there is not enough evidence either way to support the assumption of stability (and, hence, of untrainability) (Carroll, 1981; Robinson, 2012; Skehan, 2002).

However, Carroll’s assumption of stability of L2 aptitude seems unjustifiably restricted in its scope. While L2 aptitude must at some level be derived from first-order abilities, it is unlikely
that the MLAT or any other L2 aptitude test measures these in any direct manner. As discussed in Chapter 2, the work-sample tests that the MLAT employs are most likely impure measures of individual L2 aptitude abilities and most likely also measure higher-order thinking, for example metacognitive (e.g. test-wiseness) strategies. Carroll’s reductionist theories of L2 aptitude that only focus on discussing four factors of L2 aptitude in isolation seem, therefore, at odds with the reality of L2 aptitude as measured by work-sample tests.

Most significant, perhaps, is the fact that Carroll offers no concrete empirical support for his assumptions of stability and untrainability. While he claims that evidence he has seen supports his views (Carroll, 1981), this evidence is never adduced. Although much research into L2 aptitude has been conducted since the development of the MLAT, most of these studies do not explicitly investigate the question of stability of L2 aptitude. The current study aims to fill this gap by focussing specifically on the issues of stability and untrainability in L2 aptitude. In the next section, the empirical studies relevant to the opposed assumptions of stability and trainability of L2 aptitude are reviewed and discussed.

3.1.2 Empirical studies of stability of L2 aptitude

Empirical evidence generally does not support Carroll’s assumption of stability of L2 aptitude. While tests of L2 aptitude have shown themselves to be internally reliable (Pimsleur, 1966; Sasaki, 2012), there is less clear evidence for test-retest reliability (Gliksman et al., 1979; Politzer & Weiss, 1969; Granena, 2013). Results from studies that have pre- and post-tested L2 aptitude show evidence of gains in L2 aptitude test scores over time (Demuth & Smith, 1987; Ganschow, 1993; Ganschow & Sparks, 1995; Sáfár and Kormos, 2008; Sparks et. al., 1992; Sparks et. al., 1997; cf. Moskovsky, Alshahrani,
Cross-sectional data show similar support for language learning experience resulting in higher L2 aptitude test scores (Eisenstein, 1980; Grigorenko et al., 2000; Planchón & Ellis, 2014; Thompson, 2013; cf. Harley & Hart, 1997; Sawyer, 1992).

Evidence in support of Carroll’s claims is limited. The MLAT itself is only one of a very few L2 aptitude tests that is extensively validated with generally high levels of internal reliability (Sasaki, 2012). As a result, few studies seem to question the reliability of the MLAT, the assumption of stability of L2 aptitude (see section 2.2.5.8), or report on reliability of L2 aptitude testing (Sasaki, 2012). In addition, only three studies could be found that had published test-retest reliabilities for any L2 aptitude tests. Gliksman, Gardner, and Smythe (1979) pre- and post-tested 614 Canadian high school students from Grades 7 to 11 with the MLAT one year apart. They found that test-retest levels were sufficient for Grades 10 and 11, but insufficient for Grades 7 to 9. They also found a grade effect for the test, with MLAT scores increasing year by year, most likely due to maturational and educational effects. However, the authors also speculated that experience with novel test tasks influenced subsequent testing. In contrast, Politzer and Weiss (1969) report on a study of test-retest reliabilities of the Pimsleur Language Aptitude Battery (PLAB) (Pimsleur, 1966) where 440 Grade 6 and 436 Grade 8 students took the test twice ten days apart. Only a small gain of about three points was recorded across all groups, suggesting the PLAB was a reliable test of L2 aptitude (but only if one assumes that L2 aptitude is stable). Granena (2013:122) also carried out an exploratory validation study of the LLAMA test. Internal reliability was acceptable, although results were not strong ($\alpha = .66-.77$) for a sample of 76 L1 Chinese, L1 English, and L1 Spanish adult speakers. However, test-retest reliability over two years was only minimally acceptable ($r = .53-.64$, $p < .05$) for a sub-sample of 20 participants. In short, while the internal reliability and validity of L2 aptitude tests are supported by evidence,
test-retest reliabilities lack strong empirical support and more work needs to be done to establish how reliable L2 aptitude tests are. As a result, the current study includes an analysis of test-retest reliability as a means of investigating stability in L2 aptitude (see section 4.6).

In the following sections (3.2.1 and 3.2.2), both longitudinal and cross-sectional studies are discussed, presenting more evidence that L2 aptitude test scores may in fact increase over time with experience in L2 learning and L2 aptitude testing. Section 3.2.3 discusses the training studies of L2 aptitude in detail and reexamines the implications of these studies for the assumptions of stability of L2 aptitude.

**3.1.2.1 Pre- and post-test studies of L2 aptitude**

The evidential basis for Carroll’s assumption of stability is unverifiable, as he (1981) cites unpublished studies as the basis of this assumption. One study that does offer support to stability in L2 aptitude comes from Moskovsky, Alshahrani, Ratcheva, and Paolini (2015), who pre- and post-tested 56 Saudi university students learning English on L2 aptitude using an Arabic adaptation of the MLAT. They found no significant differences after seven months on any of the measures. Only the vocabulary and written code test – equivalent to the MLAT III Spelling Clues test – showed a small gain. However, it must be noted that the Arabic adaptation of the MLAT used in this study was unvalidated and so results may not be reliable.

Studies that have pre- and post-tested L2 aptitude have generally shown L2 aptitude test scores to increase with L2 learning experience. Importantly, gains were evident for all L2 learners from both treatment and control groups.
Sáfár and Kormos (2008) investigated the stability of L2 aptitude of Hungarian high school students. Participants were 40 students completing a one-year course of intensive L2 English instruction at a bilingual Hungarian high school and 21 students receiving four 45-minute classes per week at a monolingual Hungarian high school. No data were reported for previous L2 learning experience. L2 aptitude pre-testing took place at the beginning of the academic year and post-testing at the end, using the HUNLAT. Results showed that the bilingual school students’ L2 aptitude measures were significantly higher than the monolingual high school students. The authors concluded that intensive instructed language learning had a strong practice effect on L2 aptitude test scores.

Sparks and Ganschow (see Ganschow & Sparks, 1995; Sparks & Ganschow, 1993; Sparks, Ganschow, Artzer, & Patton, 1997; Sparks, Artzer, Patton, Ganschow, Miller, Hordubay, & Walsh, 1998; Sparks, Ganschow, Fluharty, & Little, 1995; Sparks, Ganschow, Pohlman, Skinner, & Artzer, 1992) have been responsible for most research bringing into question the assumption of L2 aptitude stability (see section 2.2.5.8). This line of research was inspired by the results from Demuth and Smith (1987), who commented that L2 learners with learning difficulties who had completed a novel metalinguistics course on L2 learning seemed to show noticeable improvements on their L2 aptitude test scores. Sparks and Ganschow developed a specialised course of instruction for L2 learners identified as at-risk\(^{10}\). Instruction involved a multisensory structured language (MSL) approach to explicitly teach target L2 phonology. The results showed gains in not only L2 aptitude scores but also L1 skills after one year of instruction (Ganschow & Sparks, 1995; Sparks et. al., 1992), with gains sustained even after another year of L2 learning (Sparks & Ganschow, 1993; Sparks, Ganschow, Artzer, & Patton,

\(^{10}\) Ganschow and Sparks (1995) identify these as learners who had failed or struggled in their prior L2 classes, had a known history of language difficulties, or had been diagnosed with learning disabilities – where a difference of at least one standard deviation between intelligence tests and academic achievement had been recorded (Sparks, Ganschow, Fluharty, & Little, 1995).
1997). Importantly, all participant groups, including the not-at-risk learners, showed gains in their MLAT scores, suggesting that L2 learning can lead to long-term L2 aptitude increases.

These studies directly challenge the assumption of stability of L2 aptitude, as changes over time for individuals are evident from the pre-/post-test design and sustained over the long-term. The current study seeks to replicate these findings by also adopting a pre-/post-test design, although due to operational constraints the time frame is much shorter and the testing instrument is different. In addition, the current study also includes an analysis of effect sizes to quantify any potential gains in L2 aptitude test scores (see sections 5.1.5 and 5.1.6).

3.1.2.2 Cross-sectional studies of L2 aptitude and L2 learning experience

L2 aptitude studies that have collected cross-sectional data of L2 learning experience and L2 aptitude test scores have generally shown support for a positive relationship between the two factors. Group comparisons show bilinguals\(^\text{11}\) have higher L2 aptitude scores than monolinguals (Eisenstein, 1980; Grigorenko et. al., 2000; Planchon & Ellis, 2014; Thompson, 2013) and that prior L2 experience also results in higher levels of L2 aptitude compared with naive language learners (Allen & Natelson, 1977; Carroll, 1967; Eisenstein, 1980; Planchon & Ellis, 2014). However, these findings are challenged by other studies that do not show an effect for L2 experience on L2 aptitude (Harley & Hart, 1997; Sawyer, 1992).

Eisenstein (1980) investigated the differences between early and late bilingualism in relation to L2 aptitude. Participants were 93 U.S. university students, 57 of whom had had no L2 exposure before age 10, while 36 had learned at least one L2 before age 10. All participants

---

\(^{11}\) Although the different types of bilingualism (e.g. simultaneous vs sequential) are noted, the empirical studies did not consistently report these participant characteristics. Therefore, the term ‘bilingual’ in the current study refers simply, albeit potentially problematically, to a person who speaks two languages.
were administered the short form of the MLAT, self-rated their L2 ability, and rated the importance of L2 learning. Findings indicated that L2 aptitude scores were highest for multilinguals, then bilinguals with formal L2 learning, then bilinguals without formal L2 learning, and lowest for monolinguals. Put otherwise, those with more language learning experience had higher L2 aptitude scores.

Grigorenko, Sternberg, and Ehrman (2000) investigated the validity of their novel L2 aptitude test, the CANAL-FT (see section 2.1.5) as well as its relationships with measures of intelligence and prior L2 learning experience. One hundred and fifty-eight adults from the United States of America were administered the CANAL-FT, the MLAT, tests of fluid and crystallised intelligence, and a questionnaire about prior language learning experience. Results showed that the larger the number of languages spoken, the higher the CANAL-FT scores, which are consistent with Eisenstein’s findings.

Planchon and Ellis (2014) investigated the effects of bilingualism and formal language training on language aptitude. Participants were 142 adults working at the Australian Department of Foreign Affairs and Trade, all of whom had previously taken the Defense Language Aptitude Battery (DLAB) (Petersen & Al-Haik, 1976), an L2 aptitude test derived from Carroll’s four-factor construct. Participants completed an online questionnaire about their prior L2 learning experience. Results showed that bilinguals scored higher than monolinguals and that bilinguals with formal L2 learning experience prior to testing scored higher than those without formal L2 learning.

Thompson (2013) investigated the relationship between multilingualism and L2 aptitude. Seventy-nine Brazilian L2 learners were administered the CANAL-FT (Grigorenko et al., 2000) and a background questionnaire. Results showed that multilinguals scored significantly
higher than bilinguals on the CANAL-FT. Interestingly, multilinguals with only slightly more
L2 learning experience than bilinguals still scored significantly higher.

In a comprehensive survey of 2,787 final-year U.S. university language majors, Carroll
(1967) found that individuals who had studied more than one L2 scored higher on the MLAT.
Similarly, in a comprehensive survey of 1,276 CIA L2 learners, Allen & Natelson (1977)
found that individuals with prior L2 learning significantly outscored those without prior L2
learning in the MLAT test, as well as in L2 entry and exit proficiency exams.

All the above cross-sectional studies suggest that L2 learning experience leads to increases in
L2 aptitude test scores. Importantly, as participants were only tested once, no training effect
on the test is possible, which suggests that L2 learning experience is the factor responsible for
gains in L2 aptitude. However, not all studies have shown support for an effect of L2 learning
experience on L2 aptitude test scores. Harley and Hart (1997) tested 65 Grade 11 Canadian
high school students with a mix of early and late L2 immersion learning backgrounds.
Results showed no significant differences for the MLAT V Word Pairs test (associative
memory) or Weschler’s (1972) Memory scale, Form 1 (memory for text), but later immersion
students did score significantly higher on the PLAB 4 Language analysis test (inductive
language learning). The authors concluded that, overall, earlier L2 learning does not result in
higher L2 aptitude than later L2 learning. It must be noted, however, that L2 aptitude in this
study was operationalised in an unconventional way. Each measure was taken from a separate
battery of tests with no no cogent argument made to substantiate such a decision.
Consequently, the results of such novel measurement of L2 aptitude testing require
qualification, especially considering the issues of measuring L2 aptitude discussed in Chapter
2 (see section 2.2.5.2.3).
In a different type of study, Sawyer (1992) administered the short form of the MLAT to 129 adults prior to commencing courses in south-east Asian L2 languages. Participants also completed a self-report questionnaire about their prior L2 learning experience, including age of onset, length, and intensity of formal L2 instruction and informal exposure, and self-ratings of proficiency for all previous L2s. Results showed no correlations between L2 learning experience variables and any of the MLAT scores. L2 learning experience variables also loaded onto a different factor to the MLAT tests in a principal components analysis. The author concluded that L2 learning experience and L2 aptitude were independent factors in L2 learning. However, closer inspection of the results shows that the number of L2s self-reported did show a small correlation with total MLAT scores, consistent with the studies above. It is important to note, however, that some of the self-reporting data may not have been normally distributed, making the results from the parametric testing less reliable.

Language learning experience emerges as an important variable in relation to L2 aptitude in the discussion above. In general, it seems that the more experience one has with language learning, the higher one’s L2 aptitude test scores (Allen & Natelson, 1977; Carroll, 1967; Eisenstein, 1980; Grigorenko et. al., 2000; Planchon & Ellis, 2014; Thompson, 2013). However, a key question is what exactly constitutes “language learning experience”? Its operationalization relies on self-reporting and varies noticeably across studies (cf. Planchon & Ellis, 2014 and Sawyer, 1992). Nevertheless, across most studies, the most important dimensions of language learning experience in relation to L2 aptitude appear to be a) the number of languages learned; b) the proficiency achieved in each language, as well as c) the nature of the learning (instructed vs naturalistic). In the current study, a measure of language learning experience was collected by asking participants to report if they spoke any additional
languages and, if so, to name those languages (many participants also self-reported proficiency levels; see section 4.2).

In summary, language learning experience appears to be an important variable in L2 aptitude test scores, with more language learning experience correlating with higher test scores. However, the term ‘language learning experience’ lacks a clear or comprehensive definition and its operationalization is inconsistent across studies, making generalizations across studies challenging. The current study included a measure of this variable based on the results of previous studies to determine what effect language learning experience has on predicting L2 aptitude test scores (see section 5.1.3). Evidently, the implication of the findings from these studies on language learning experience is, that the process of learning a language has a training effect on L2 aptitude (Sáfár & Kormos, 2008) that exceeds any test training effect measured between a pre- and post-test. Undoubtedly, the issue of trainability of L2 aptitude is not well understood and has been under-researched (see section 2.2.5.8). We now turn to the small number of studies that have been conducted on the trainability of L2 aptitude.

3.1.2.3 Training studies of L2 aptitude

There are only two studies that explicitly investigated the trainability of L2 aptitude abilities. While both are suggestive of the potential for training L2 aptitude, each illustrates the difficulties in achieving clear results from such a complex learning situation with many interrelated variables.

In a two-phase field experiment, Politzer and Weiss (1969) attempted to improve L2 achievement by explicit training in L2 aptitude. Each phase employed a pre-test, treatment-control, post-test design. In reality, four cohorts of participants were studied. Phase
I of the study included 247 adults undertaking intensive L2 courses at the U.S. Defense Language Institute (DLI) and 356 high school students undertaking regular foreign language courses as part of their school work. Phase II included 268 high school students undertaking normal L2 courses as part of their school work and another 32 high school students not taking any L2 studies (non-L2). All participants, except the non-L2 high school students, were taking a variety of L2s. The experimental groups (approximately half of the participants) in phase I and II were given L2 aptitude training materials in addition to their L2 learning. Control groups in phase I did not undertake any additional L2 learning in place of the aptitude training, while control groups in phase II underwent cultural studies relevant to their target L2 for an equivalent amount of time to the training undertaken by the experimental groups. Pre-testing and post-testing took place at the beginning and end of the academic year and included both L2 aptitude tests and specially created L2 proficiency measures. The DLI participants were tested using the Army Language Aptitude Test (ALAT); the phase I high school students were tested using the MLAT II, III, and IV; and the phase II high school students were tested using the PLAB 3, 4, 5, and 6. Training in Phase I was based on the MLAT and included learning metalinguistic content about a variety of languages not necessarily related to the L2s participants were learning. Phase II training was more skills-based and placed more emphasis on inductive language learning. No statistically significant differences were found between any of the groups in either of the phases, although the general trend was for the experimental groups to score more highly than the control groups. This result is generally accepted in the literature as evidence that L2 aptitude is untrainable (Carroll, 1981). However, one interesting finding did emerge. In phase II of the study, a test-retest control group that did not study any L2 or undergo any L2 aptitude training completed a pre- and post-test of the PLAB eight weeks apart. The mean gain score for the
post-test was approximately 3 points. Almost all experimental and control groups except one achieved gains above this test-retest effect, including those students who received the aptitude training but were not undertaking any L2 learning. The authors concluded that L2 learning seems to increase L2 aptitude scores above the test-retest level, but that L2 aptitude training was best given prior to L2 learning. They also identified the difficulty in clearly separating out the effects of L2 learning and aptitude training from post-test gains in L2 aptitude testing. Similar results were found in the current study (see section 5.2).

While these results are not conclusive, they are definitely suggestive, especially considering the complexity of the study design. In particular, three main problems are important to note when interpreting the results of the Politzer and Weiss (1969) study. First, the efficacy of the training implemented was unclear in respect of both its content and delivery. The L2 material included in the training was not in the target L2 of the participants, which may have made it more difficult to understand and less engaging for participants. Aptitude training tasks were also done for only approximately 10 minutes, two to three times a week during normal classes. Therefore, participants may not have had enough time to benefit from the training. Second, feedback from schools revealed a strong negative attitude towards the training, suggesting that even those participants who completed the materials may not have actively engaged with them. In fact, the authors acknowledged that many classes openly refused to complete all of the training, citing that they did not see it as beneficial to their L2 learning. Third, control groups also improved in L2 aptitude scores, but the study was unable to distinguish between the effects of L2 learning and the training. These three limitations suggest that the weak results of the study are quite possibly unrelated to the question of
untrainability of L2 aptitude and are more likely the result of the training design and implementation, as well as the statistical analysis of the results.

In the only other explicit study of L2 aptitude training, Yeni-Komshian (1965) trained high school students in phonological abilities prior to commencing an L2 course. Yeni-Komshian’s PhD study involved pre- and post-testing high school students on a range of phonological aptitude measures. Experimental groups received two two-hour training sessions in which they undertook practice exercises in phonological abilities. Control groups completed only the pre- and post-tests. Results showed that the training was effective in increasing aptitude test scores but primarily for those who were already of a ‘high’ level (i.e., those who had achieved approximately 80% and above in the pre-test). Importantly, the whole study and the training were extremely short and limited. Gains measured could have been short-lived effects – which is in line with Carroll’s (1981) hypothesis – and other students with more training could have achieved higher test scores. In addition, participants below the ‘high’ initial level of aptitude did not benefit from the training, which did not seem to include any explicit explanation of tasks. Therefore, the content may have been appropriate for all, but delivered in such a way that only apt learners could benefit. Despite these limitations, the overall conclusion of this study supports the trainability of L2 aptitude.

In summary, both studies of L2 aptitude training suggest that L2 aptitude abilities are trainable. However, neither study is conclusive, and a range of factors must be taken into account before regarding them as evidence for trainability. First, the efficacy of a training course needs to be established. A lack of significant gains in L2 aptitude scores is not necessarily evidence that L2 aptitude is untrainable, but could instead indicate that a training course was ineffective. Neither is it the case that significant gains in L2 aptitude scores are
necessarily evidence that L2 aptitude is explicitly trainable, particularly if training takes place in conjunction with L2 learning. Second, both previous training studies may have been confounded by maturational effects, as participants were teenagers and not adults. Third, the use of an experimental-control group design also fails to account for the roles of L2 learning and the efficacy of explicit training separately. Instead, statistical models that look at the explanatory or predictive strength of training may offer more insight into the efficacy of any training course on L2 aptitude scores. Finally, the learning design of any training materials needs to be taken into account. Time on task may also be an important factor in the efficacy of training. How much time in total, as well as how often and over what period of time training is conducted, may all prove to be important variables in the efficacy of developing L2 aptitude abilities.

In short, much more longitudinal research on the trainability of L2 aptitude is needed to investigate whether L2 aptitude scores can be enhanced through training, and if so, which types of training produce which types of effects. The current study seeks to address the problems of these previous studies by (1) opting for a regression design with variables from the training course predicting L2 aptitude post-test scores; (2) recruiting adults\textsuperscript{12} from a university L2 learner population to avoid any maturational effects; (3) using a random forest regression model capable of analysing a wide range of variables and judging their ability to predict L2 aptitude post-test scores; and (4) collecting data on how long and how regularly participants spent on training tasks and including these variables as predictors of post-test L2 aptitude test scores (see chapter 4 for more details).

\textsuperscript{12} The term adults here is used in contradistinction to high school teenagers, although some participants were still technically teenagers (18-20 y.o.).
3.1.2.4 Summary of evidence on purported stability and untrainability of L2 aptitude

Overall, the data tend to suggest that L2 aptitude is not as stable as traditionally assumed and, correlatively, that particular types of training delivered with appropriate methods and at particular times may effectively enhance L2 aptitude. This evidence come from both longitudinal and cross-sectional data, from a variety of learners in different contexts. However, there are issues in the research yet to be resolved. L2 aptitude testing may be susceptible to maturational effects, challenging the results of studies conducted with high-school aged participants. Longitudinal studies seem more important to future research, as these last might avoid any issues with self-reporting that may influence cross-sectional data. Finally, trainability studies offer many potential insights into the question of stability, although any L2 aptitude training material needs to be targeted and relevant to participants and to be delivered as effectively as possible; in short, to be as engaging and motivating as possible. The current study seeks to address these issues by developing training materials that are more relevant for L2 learners, and by collecting data on a range of variables that may explain post-test scores.

3.1.3 Alternatives to Carroll’s perspective of L2 aptitude

By identifying theoretical limitations and challenging the empirical evidence, the previous sections have cast doubt upon the assumptions of stability and untrainability in L2 aptitude first advanced by Carroll. It is unsurprising, then, that despite Carroll’s tremendous influence, his contributions to L2 aptitude testing and theory have not been universally accepted by researchers in the field (e.g. Neufeld, 1978). Different models of L2 aptitude seek to situate and explain L2 aptitude within particular alternative frameworks. For example, the Linguistic
Coding Deficit Hypothesis (Sparks, 1995) explains the link between L2 aptitude and L1 skills; similarly, Gardner’s (1985) socio-cultural model explains the role of L2 aptitude in relation to motivational and attitudinal factors in predicting L2 achievement (Garcia, 1996). However, these alternative views do not account for why L2 aptitude test scores may change over time. Only perspectives of L2 aptitude from (general) cognitive psychology offer such an explanation. In particular, McLaughlin’s (1995) information processing view of L2 aptitude, Sternberg’s (2002) model of L2 aptitude, and the ATI perspective of L2 aptitude offer the clearest possible explanations of how L2 aptitude is sensitive to experience and therefore changes over time. While these theories of L2 aptitude have not been extensively researched or validated, they do offer possible explanations for the evidence of L2 aptitude test scores that improve commensurate with experience in language learning. Although these theories differ in their details, they agree in important respects that I will discuss in what follows.

Cognitive theories of L2 aptitude differ from Carroll’s traditional construct in four important ways. First, while Carroll (1962) assumes that L2 aptitude is language-specific, cognitive theories view L2 aptitude as domain-general, construing language as a complex skill that is learned in the same way as any other (McLaughlin, 1995; Sternberg, 2002). Second, while Carroll’s view of L2 aptitude seems limited to first-order abilities (see section 3.1), cognitive perspectives of L2 aptitude are more holistic (Sternberg, 2002; Snow, 1991) and hierarchical (McLaughlin, 1995; Robinson, 2001). Third, while Carroll assumes that L2 aptitude is a stable system, cognitive perspectives view L2 learners as adaptive, deploying learning systems that evolve in direct response to experiences in language learning (McLaughlin,

---

111

13 Snow’s (1991) work on aptitude-treatment interactions is not specific to L2 aptitude, but is applicable to aptitude theory in general and has been pioneered in L2 aptitude research by Robinson (2001, 2005; see chapter 1.1.4). However, this discussion will focus more on Snow’s contributions rather than Robinson’s, as Robinson’s model is more developed and supported by empirical evidence.
1995; Sternberg, 2002; see also Snow, 1991). Fourth and finally, cognitive perspectives of L2 aptitude predict that strategies and (meta)knowledge of L2 learning can be taught (McLaughlin, 1995; Sternberg, 2002). These four points are important to the current study, as they provide the theoretical basis for investigating stability of L2 aptitude, and are central to the design of the study’s training treatment. Each of these four points is discussed in more detail below.

3.1.3.1 Second language aptitude is based on general cognitive abilities

All three cognitive perspectives of L2 aptitude view language learning as a complex cognitive skill guided and regulated by concepts and mental systems. Language learning is viewed the same as any other type of complex skill, although in the case of language, processing is specific to the linguistic domains of phonology, morphology, semantics, syntax, and pragmatics (McLaughlin, 1995; Sternberg, 2002). However, each perspective takes a slightly different approach. The aptitude-treatment interactionist perspective is by the far the most general and broad ranging. Snow & Lohman (1984:347) define aptitude as the situation when “inner” and “outer” environments align and a goal is reached with little or no impediment, taking into account all individual differences, e.g. motivation, not just cognitive skills. Sternberg’s (2002) theory of L2 aptitude is compatible in most respects with ATI perspectives, although it is based on Sternberg’s (2003) own theory of triarchic/successful intelligence, which consists of analytical, creative, and practical abilities. Sternberg’s theories of L2 aptitude have been tested through the CANAL-FT L2 aptitude test, which views L2 aptitude as built primarily on the general cognitive ability to deal with novelty and ambiguity in the linguistic domain (Grigorenko et al., 2000; see section 2.1.5). McLaughlin (1995)
views L2 learning from an information-processing perspective, in which the acquisition of skills involves an initial phase of controlled processing that requires large amounts of attention resources. Through practice, skills become routinised and stored in long-term memory, which requires fewer attention resources. On this view, one part of L2 aptitude would be the automatisation of sub-skills of L2 learning such as phonemic coding, pattern recognition, or associative memory. In short, while the three perspectives of L2 aptitude differ in their specifics, they all propose models in which general cognitive abilities are applied to the specific tasks of language learning. This generalised view was adopted in the design and creation of the current study’s language analytic abilities training (see section 3.2.3), and provides the theoretical basis for explaining its results (see chapter 6).

3.1.3.2 Second language aptitude as a hierarchical system

All three cognitive perspectives view L2 aptitude as hierarchical (McLaughlin, 1995; Robinson, 2001, 2005; Sternberg, 2002). At the most basic level, Snow and Lohman (1984) and Sternberg (2002) identify a performance variable, which would equate to the ability to execute one of McLaughlin’s (1995) sub-skills of L2 learning, e.g. pattern recognition. This would also appear to equate with Carroll’s (1962) four first-order abilities of phonemic coding, associative memory, grammatical sensitivity, and inductive language learning. Snow and Lohman (1984) identify these performance variables as reflections of elementary, or first-order, components that are relatively stable, consistent with Carroll’s assumptions of stability for his construct of L2 aptitude. Importantly, however, all three cognitive perspectives view L2 aptitude as consisting of more than just these first-order abilities.
Sternberg (2002) and McLaughlin (1995) both discuss the strategic deployment of first-order abilities to a given L2 learning task. A simple example of this would be the acquisition of L2 vocabulary: one strategy might be to learn L2 words individually with a direct L1 translation; another strategy might be to learn L2 words in a phrase, which might also include pragmatic and syntactic information, as well as a semantically and pragmatically equivalent L1 translation of that phrase. Both approaches are strategically different while solving the same L2 learning problem. Snow and Lohman (1984) also identify this use of strategies, but instead term it sequencing – the specific application of related sub-skills to the task – to factor in how individuals might vary in their sequenced application of a strategy’s sub-skills to the same task over time. Using the previous example, a single L2 learner might use both strategies for acquiring L2 vocabulary at different points in time, and even swap back and forth between approaches. These strategies in L2 learning involve metacognitive skills of planning, monitoring, and evaluating task performance (McLaughlin, 1995; Sternberg, 2002). McLaughlin (1995) explains how the elaboration of sub-skills into complex skills deployed strategically involves restructuring, loosely defined as a qualitative change from one stage of development to the next (McLaughlin 1995:374). McLaughlin (1995) cites two studies (Nayak et al., 1990; Nation & McLaughlin, 1986) that illustrate the metacognitive knowledge of L2 learners. In these two studies, experts differed from novices in their ability to more flexibly switch between strategies. McLaughlin (1995:) claims this reflects an increased metacognitive ability to evaluate the effectiveness of strategies being employed and to consequently restructure their concepts of L2 grammatical systems. This view of L2 aptitude as a suite of metacognitive skills, chief among which is a restructuring capacity, concurs with research by Cummins & Gulutsan (1975) that found participants’ ability to mentally switch between linguistic systems correlated with the MLAT III and MLAT IV sub-tests.
McLaughlin defines this metacognitive component of L2 aptitude as the ability to flexibly “restructure and reconfigure linguistic representations” in working memory (McLaughlin 1995:383), a view which is plainly compatible with the construal of working memory as L2 aptitude (Miyake & Friedman, 1998).

Above both of these levels of aptitude are knowledge acquisition components, which involve learning how to solve the problem at hand or acquire declarative knowledge (Sternberg, 2002). Specifically, “experts replace complex sub-elements with single schemata that allow more abstract processing” (McLaughlin 1995:378) while novices “focus principally on the surface elements of the task” (ibid.). On this account, experts have more insight into the nature of the L2 learning problems they face and can execute a series of planned, interconnected actions in response. Novices, on the other hand, may lack this understanding and consequently deal with each problem at face value without seeing its connection to other problems. In L2 aptitude testing, Snow and Lohman (1984) suggest that this understanding of the entire learning process emerges not as scores in an individual sub-test, but rather as a summation across the whole test. In other words, no individual question or sub-test measures this knowledge of L2 learning, but aggregate scores of individuals may reflect differing levels of such knowledge. These models of L2 aptitude as involving both strategic and metacognitive knowledge of L2 learning explain why work-sample tasks might be valid predictors of L2 learning – they measure all three levels of L2 aptitude. In particular, a hierarchical view of L2 aptitude might also help to explain why simple measures of WM, e.g. phonological short-term memory (PSTM), are less predictive of L2 achievement than complex measures of WM, e.g. reading span tasks (Li, 2015; see section 2.2.5.2.2).
These theories suggesting that improvements in L2 aptitude are unlikely to be due to improvements in first-order sub-skill performance, but are rather likely due to changes in strategy or insight into the L2 learning process as a whole, have heavily influenced the current study. These insights determined that the training treatment of the current study focus on instructing participants in L2 learning strategies and about processes and stages in L2 learning, rather than on intensive practice of first-order sub-skills. In addition, to capture possible contributions of summation effects, aggregate L2 aptitude test scores were included as a variable in predicting post-test scores (see section 5.2.2).

3.1.3.3 Second language learners as adaptive

All three cognitive theories of L2 aptitude also view learners as adaptive (McLaughlin, 1995; Sternberg, 2002; Snow, 1991). Sternberg’s (2002:15) general view of L2 aptitude is apparent in his noting that:

One’s ability to achieve success depends on one's capitalizing on one's strengths and correcting or compensating for one's weaknesses through a balance of analytical, creative, and practical abilities in order to adapt to, shape, and select environments. People have different patterns of abilities, and they will learn a language successfully when the way they are taught fits their ability patterns.

On this view of L2 aptitude, individuals adapt to the demands of learning another language by, for example, developing new skills or strengthening existing skills, which may include improving weak ones. Individuals may also reconceive of and shape their L2 learning tasks to better suit their approaches by, for example, repurposing analytical tasks like grammar translation into memorisation tasks. Finally, individuals may also select and/or completely
change their L2 learning task by, for example, deciding to enrol in a formal L2 course as opposed to self-study.

Similar to Sternberg’s view, Snow (1991; Snow & Lohman, 1984) defines aptitude as a dynamic cognitive system in which learners respond to experience and adapt. This adaptation takes place both in the short and long term. For example, learner approaches to questions over the course of the entire test can evolve, an adaptation which is evidenced in the total test scores. Snow and Lohman (1984) claim that in a dynamic cognitive system, assembly and control processes are the most important factor underpinning performance, sequencing, and summation processes. Assembly and control processes include the initial organisation of sequencing performance processes, and the subsequent reorganisation of these processes after monitoring their effectiveness. In this way, assembly and control processes are higher-order processes that underpin learner adaptation to a task. This view of the learner is also consistent with the studies cited in the previous section (see Cummins & Gulutsan, 1975; Nayak et al., 1990; Nation & McLaughlin, 1986).

According to Sternberg (2002), individuals gain expertise with experience and feedback, thereby becoming more effective learners for the next time they are confronted with the same or an analogous learning situation. On this view, L2 aptitude is construed as dynamic and sensitive to experience. Sternberg’s theory of L2 aptitude is operationalised in the Cognitive Ability for Novelty in Acquisition of Language (Foreign) Test (CANAL-FT) (Grigorenko et al., 2000) (see section 2.1.5 for details), a dynamic test that measures an individual’s L2 learning performance over time with feedback, rather than a static, once-off test, like the MLAT. If L2 aptitude is the ability to adapt to the outer environment, then this would suggest that dynamic testing such as that used by the CANAL-FT (also see Krug, Shafer, Dardick,
Magalis, & Parente, 2002) has more to offer L2 aptitude research. These understandings of
the learner as adaptive and of L2 aptitude as a dynamic system sensitive to experience are
central to the current study, underpinning the hypothesis that L2 aptitude is not stable and is,
therefore, trainable.

3.1.3.4 Strategies and knowledge of second language learning can be taught

Finally, all three cognitive theories of L2 aptitude regard the strategies and knowledge of
second language learning as teachable (McLaughlin, 1995; Snow & Lohman, 1984;
Sternberg, 2002). McLaughlin (1995) discusses studies in which multilingual individuals
demonstrated greater flexibility in applying L2 learning strategies, and refers to studies that
have attempted to teach such strategies to “poor” learners, although he does not discuss their
success. Sternberg (2002) cites various studies in non-L2 learning in which strategies in
creative, analytic, and practical abilities were taught to less-experienced learners with
significant gains in post-testing recorded. There has also been some work done on training
working memory, which has been claimed to be central to L2 aptitude (Miyake & Friedman,
1998) and has been a strong area of interest in L2 aptitude research recently (see section
2.2.5.2.2). Using a dynamic training task, Jaeggi, Buschkuehl, Jonides, and Perrig (2008)
found that increases in working memory performance from the training tasks were
transferable to measures of fluid intelligence, a related but independent construct.
Interestingly, they also found that the amount of time spent on training correlated with the
amount of gains achieved between pre- and post-test. This suggests that while training can be
effective in increasing aptitudinal abilities, the amount of training completed is an important
variable. For this reason, the current study included variables for the amount of time on task,
as well as for how much of the training treatment was completed (see section 5.2.1 for details).

Snow and Lohman (1984) review a series of studies that investigated the efficacy of teaching learning strategies and ability test strategies. Overall, these studies found that teaching these strategies correlated with increases of post-test scores, but predominantly for individuals with initially low test scores. Individuals with initially high test scores actually experienced decreases in post-test scores, and claimed that the strategies they were taught interfered with the strategies they normally used. This classic case of aptitude-treatment interaction effects highlights an interesting caveat to the idea that the strategies and knowledge of learning are teachable – not all individuals will benefit equally from the same strategies or knowledge of learning. The current study factored this finding into the design of its training. As part of the instructional treatments, participants were explicitly told that the strategies and ideas in the training course would not always be suitable for them, and were urged to experiment in their own language learning with the strategies and ideas presented, to reflect on the efficacy of these strategies, and to use those that helped them in their L2 learning and discard those that did not (see section 4.4.4.1). The analysis of the data also investigates any possible ATI effects in L2 aptitude stability by examining gain scores for participants with below average and above average pre-test scores for all sub-tests (see section 5.1.6).

3.1.3.5 Summary of alternatives to Carroll’s perspective of L2 aptitude

In summary, general cognitive perspectives of L2 aptitude challenge Carroll’s assumption of stability of L2 aptitude without contradicting his claim that first-order abilities of L2 aptitude are limited and stable over the long term. Evidently, the key difference is one of scope.
Carroll appears to regard his four factors of L2 aptitude (phonemic coding, associative memory, grammatical sensitivity, and inductive language learning) as first-order abilities with neurobiological limitations. The L2 aptitude theories of Sternberg (2002), McLaughlin (1995), and aptitude-treatment interactionist perspectives expand upon Carroll’s view by including the higher order components of L2 aptitude strategic and metacognitive language learning knowledge. These higher-order components recognise the agency of the L2 learner (Sasaki, 1993) and allow a role for experience in language learning to create changes in the L2 aptitude system without necessitating improvements in first-order sub-skills that can be considered relatively stable (Snow & Lohman, 1984). As a result, the learner is understood as an agent who plans, monitors, self-evaluates on task performance and adapts strategies and approaches accordingly (McLaughlin, 1995; Sternberg, 2002). L2 aptitude strategies and knowledge of L2 learning are considered to be teachable, although the evidence from analogous studies in other fields of learning suggests that instruction in these areas will benefit individuals with low aptitude levels more than those with high aptitude levels (Snow & Lohman, 1984). In addition, the effectiveness of instructional treatments may depend on the amount of training undertaken (Jaeggi et al., 2008). The current study has adopted these theoretical perspectives on L2 aptitude to hypothesise that L2 aptitude test scores will increase with L2 learning experience and aptitude instruction, although this increase will be dependent upon initial L2 aptitude levels and the amount of time spent on training.

3.1.4 Summary of stability and trainability of L2 aptitude

While certain theoretical perspectives and highly suggestive empirical data support the proposition that L2 aptitude is unstable and, therefore, trainable, the contentious issue of the
stability or instability of L2 aptitude lacks a definitive answer. Although a range of empirical evidence suggests that L2 aptitude is sensitive to experience and is enhanced through L2 learning and even training in L2 aptitude abilities, very few studies have investigated these issues explicitly with no L2 aptitude training studies in over 40 years. It should also be noted that L2 learning experience is used broadly to encompass all events in L2 learning, including but not limited to L2 instruction, L2 aptitude training, and naturalistic L2 learning situations. In addition, L2 aptitude research is centred around L2 aptitude tests, of which the MLAT is considered the benchmark. Consequently, Carroll remains a central figure in L2 aptitude theory, a field of inquiry in which much of the research stems from Carroll’s groundbreaking work in developing the MLAT. However, Carroll’s conclusions regarding the stability of L2 aptitude are at odds with much of the data and lack any explanatory power. By contrast, more recent cognitive theories of L2 aptitude do seem to account for increases in L2 aptitude measures over time. This explanatory power stems from having a more holistic view of L2 aptitude, which extends beyond first-order abilities to encompass higher-order thinking. This broader perspective also seems to more accurately characterise the work-sample tasks that typify L2 aptitude testing. Nevertheless, more research that directly examines the purported stability and potential trainability of L2 aptitude is needed to clarify these issues. The current study seeks to investigate this stability and its correlative, untrainability, by pre- and post-testing all components of L2 aptitude and investigating the efficacy of a training course intervention targeting only one of the components of L2 aptitude, namely, language analytic abilities. The following section of this chapter discusses language analytic abilities in the context of a training course.
3.2 Language analytic abilities: a training perspective

Despite broad agreement as to the componential structure of L2 aptitude, most research has only considered the construct as a whole (see section 2.2.5.5). Arguably, the most under-studied of the components of L2 aptitude, particularly in relation to stability and trainability, is that of language analysis (Skehan, 2002). Treatments focussing on phonological abilities have been studied (see section 3.1.2.1 for discussion of Sparks & Ganschow’s work and section 3.1.2.3 for Yeni-Komshian, 1965), as has the role of (working) memory in L2 aptitude (see Chen, 2013; Erlam, 2005; Fujii, 2005; Howes, 1999; Robinson, 2005; Roehr & Gánem-Gutiérrez, 2009; Serafini, 2013; Suzuki & DeKeyser, 2016; Tolentino & Tokowicz, 2014; Vandergrift & Baker, 2015; Winke, 2013; Wistner, 2014; Yalçın et al., 2016; Yilmaz, 2013). In contrast, no studies have focussed solely on treatments of language analytic abilities, despite the fact that such studies might crucially advance understanding of L2 aptitude as a whole. Notwithstanding this relative lack of research, language analytic abilities (LAAs) are still considered to be a strong predictor of L2 achievement overall (see Li’s 2015 meta-analysis). Plainly, more research is required to understand what exactly constitutes LAAs and how sensitive to experience and training they may or may not be. This chapter discusses these issues by first reviewing definitions (section 3.2.1) and examining what we currently know about LAAs (section 3.2.2). The following section proposes a reconceptualisation of LAAs (section 3.2.3), including a list of potential abilities (section 3.2.4). Finally, this new conceptualisation of LAAs is discussed in relation to its implications for the trainability of LAAs and the training regime implemented in the current study (section 3.2.5).
3.2.1 Terms, definitions, and issues with language analytic abilities

Skehan (1998, 2002) attempted to update the traditional L2 aptitude construct by combining two of Carroll’s original four factors of L2 aptitude into one component with a single descriptor, language analytic abilities. Language analytic abilities subsume grammatical sensitivity and inductive language learning ability, which Skehan argues really refer respectively to the passive and active dimensions of the same factor.

Grammatical sensitivity allows concentration on one word (though in a larger structure), and only requires the test-taker to recognize, in whatever manner, the function the word fulfils so that it can be matched with another word. Inductive language learning operates upon longer structures and involves manipulation of a pattern at a greater degree of consciousness. But both sub-tests try to sample within the area of a general analytic ability, simply doing so in different ways, which, of course, could lead to slightly different patterns of correlation (Skehan 1998:201).

Two important points are worth noting here. First, like L2 aptitude in general, language analytic abilities are defined in terms of tests that purport to measure the construct (see section 2.1). In other words, language analytic abilities are what tests of language analytic abilities measure. Second, Skehan seems to claim that language analytic abilities are a subset of general cognitive analytic abilities. Put otherwise, language analytic abilities are non-specific to L2 learning and are a subset of domain-general abilities applied to the specific context of L2 learning, consistent with the views of McLaughlin (1995) and Sternberg (2002). Both of these points are discussed in more detail in the next section, in which exemplars of language analytic abilities tests are also described and relevant empirical studies reviewed for what they reveal about the construct.
3.2.2 What we know about language analytic abilities: tests and empirical studies

Language analytic abilities are understood to be what tests of language analytic abilities measure. Most of what we know about the construct is inferred and/or deduced from the test tasks themselves and the statistical relationships that these tests share with other variables and constructs relevant to L2 learning. Consequently, many of the criticisms that are true for L2 aptitude in general (see section 2.1 for discussion) also hold true for language analytic abilities. In this section, we first look at the tests of language analytic abilities, their theoretical constructs, and questions that remain unanswered. Next, the empirical research related to language analytic abilities is reviewed to gain further insights into the construct.

3.2.2.1 Measures of language analytic abilities

In the absence of a robust definition of language analytic abilities, the test tasks that purportedly measure this construct offer us some insight into how LAAs are understood. Arguably, the MLAT IV Words in sentences sub-test, the PLAB 4 Language analysis sub-test, and the LLAMA F test are the three most used tests of language analytic abilities. The three distinct tasks employed in these three tests can also be considered to exemplify language analysis in L2 learning. Table 3.1 summarizes what each test purportedly measures and the task types employed.
### Table 3.1 Descriptions of the MLAT IV, PLAB 4, and LLAMA F measures of language analytic abilities

<table>
<thead>
<tr>
<th>Test</th>
<th>What it measures</th>
<th>Task type</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLAT IV Words in sentences</td>
<td>Grammatical sensitivity (Carroll, 1962)</td>
<td>Identify the function of words and phrases from analogous examples in different sentences (Carroll, 1962)</td>
</tr>
<tr>
<td>PLAB 4 Language analysis</td>
<td>Inductive language learning (Carroll, 1973) / linguistic reasoning (Pimsleur et al., 1963)</td>
<td>Figure out how to translate statements from English into an artificial language (Pimsleur, 1966)</td>
</tr>
<tr>
<td>LLAMA F</td>
<td>Grammatical inference (Meara, 2005) / explicit inductive language learning (Granena, 2013)</td>
<td>Work out the grammatical rules of an artificial language from a set of 20 pictures and L2 sentences that describe them (Rogers et al., 2017)</td>
</tr>
</tbody>
</table>

In the MLAT IV Words in sentences test, two sentences are given. In the first sentence, a key word/phrase is highlighted using capital letters and underlining. In the second sentence five words/phrases are underlined. The task is to select one of the five given words/phrases in the second sentence that has the same grammatical function as the highlighted word/phrase given in the first sentence (see figure 3.1 for an example). Carroll (1962) labels the MLAT IV a test of grammatical sensitivity and as “a grammatical analogies test in which, for each item, the examinee is asked to find a grammatical construction, in one or more English sentences, that has a function analogous to that of an indicated word or phrase in a key sentence” (Carroll 1991:19). Performance on the MLAT IV may be improved by metalinguistic knowledge, although this is not always the case (Carroll, 1962). Sáfár & Kormos (2008) argue that when the grammatical function of words in sentences is explicitly taught, e.g. in school curricula of L1 grammar, then the MLAT IV is actually a measure of literacy skills (Detschelt, 1993; Sparks et al., 2011) and not of cognitive abilities. However, the relationship between
grammatical sensitivity and metalinguistic knowledge might not be one way. Grammatical sensitivity might possibly build metalinguistic knowledge by helping the learner to categorise words by grammatical function (Roehr, 2007). In summary, the MLAT IV Words in sentences test of grammatical sensitivity possibly measures both grammatical knowledge, be it implicit or explicit, and/or linguistic analogical reasoning. Again, the issue here is that the MLAT IV is an impure measure of language analytic abilities in which the role played by each of these factors in L2 learning is not known.

In each of the following questions, we will call the first sentence the key sentence. One word in the key sentence will be underlined and printed in capital letters. Your task is to select the letter of the word in the second sentence that plays the same role in that sentence as the underlined word in the key sentence.

Look at the following sample question:

Sample: JOHN took a long walk in the woods.

Children in blue jeans were singing and dancing in the park.

A                        B                 C                  D                   E

Figure 3.1 Example question from the MLAT IV Words in Sentences sub-test (LLTF, 2013)

The PLAB 4 Language analysis test consists of a list of two nouns, two sentences, and one conjugated verb given in an unknown L2, with English translations provided. A series of English sentences with four possible translations in the L2 are given based on the information in the list. The task is to infer the translation from English into the L2, and then select the answer closest to your own (see figure 3.2). The PLAB 4 is viewed as one of the best tests of inductive language learning (Carroll, 1973) and is claimed to test verbal reasoning (Pimsleur et al., 1963). The test purportedly measures “the ability to examine language material (in either auditory or printed form) and from this to notice and identify patterns of correspondences and relationships involving either meaning or grammatical form” (Carroll 1973:5-6). Since the PLAB was designed primarily for high school students, it is possible that
the sub-tests may not be difficult enough to discriminate between adult L2 learners with higher levels of ability. In summary, inductive language learning may involve noticing, induction, and verbal reasoning of linguistic input to extrapolate L2 knowledge. Similar to the MLAT IV, the PLAB 4 is an impure measure of a single factor and the roles that each ability of analysis plays in L2 learning remains unknown.

The list below contains words from a foreign language and the English equivalents of these words.

<table>
<thead>
<tr>
<th>jiban</th>
<th>boy, a boy</th>
</tr>
</thead>
<tbody>
<tr>
<td>jojo</td>
<td>dog, a dog</td>
</tr>
<tr>
<td>jiban njojo za</td>
<td>A boy likes a dog</td>
</tr>
</tbody>
</table>

By referring to the above list, figure out how the following statement should be expressed in this language. Do this without writing on paper.

A dog likes a boy.

Figure 3.2 Example question from the PLAB 4 Language analysis sub-test (LLTF, 2013)

The LLAMA F test, by contrast, consists of two parts. In the first part, 20 phrases are presented with associated images. The examinee is given five minutes to learn the underlying L2 system. In the second part, novel images are presented with two possible answers. The task is to identify the underlying grammatical patterns evident from the learning stage to correctly select the sentence that matches the novel image in the testing stage (see figure 3.3). The test aims to measure the ability to analyse linguistic data from the bottom up and infer underlying grammatical patterns and structure. It is also a test of explicit learning (Granena, 2013) and is related to or involves WM to some extent (Yalçın et al., 2016). The LLAMA F test appears to measure all three constructs of grammatical inference (as part of a language analytic abilities), explicit L2 aptitude, and WM.
While other measures of language analytic abilities do exist, e.g. CANALT-FT Section 4: Sentential Inference and Section 5: Learning Language Rules, their use has not been widespread and so their contribution to the construct has been limited. However, it is worth noting Skehan’s (1986) test of metalinguistic rule rehearsal, in which learners are presented with a list of grammatical rules for Bahasa Indonesia. Time was given to memorise the rules for later recall. While no measure of production was included in this test, the ability to recall L2 grammatical rules is implied in any form of top-down analysis of L2 input and so this test may tap into this type of language analysis.

Overall, the MLAT IV, PLAB 4, and the LLAMA F aim to measure an ability by which learners are “analyzing given material to extract function… to make generalizations and extrapolate.” (Skehan 2002:71). In particular, the MLAT IV and the PLAB 4 have been extensively validated as part of the MLAT (Carroll & Sapon, 1959) and the PLAB (Pimsleur, 1966) respectively. The current study, therefore, used these tests as the basis for designing
new tasks of LAAs for its training course (see section 4.4.3) and the LLAMA F as a measure of LAAs for reasons that will be explained in section 3.4.1. However, the MLAT IV, PLAB 4, and LLAMA F test tasks also differ substantially, each work-sample test including measures of other abilities. The MLAT IV appears to involve metalinguistic knowledge and possibly L1 literacy skills; the PLAB 4 may require general verbal reasoning; and the LLAMA F appears to measure an aptitude for explicit L2 learning and require some (unquantified) level of WM to complete the task. These factors have also been highlighted in the small number of empirical studies relevant to language analytic abilities, a review of which is conducted in the following section.

In summary, the breakdown of the test tasks suggests that language analytic abilities in fact comprise a range of sub-abilities, each contributing to L2 learning in some way, although the nature and extent of these contributions are not precisely known. The current study seeks to investigate this issue further by providing a more detailed explanation of language analytic abilities (see section 3.2.3 and 3.2.4). This more detailed explication of LAAs was used to create a training course which included a range of tasks designed to analyse L2 input in different ways (see section 4.4.3)

### 3.2.2.2 Studies of language analytic abilities

An analysis of the empirical research into language analytic abilities provides us with insights into a range of questions raised in L2 aptitude research. Apart from the obvious relationship between measures of LAAs and L2 achievement, research has also investigated the effects of LAAs under different types of learning situations, the relationship between LAAs and

---

14 In the current study, participants were permitted to take notes, which may have reduced any potential WM load in the language analysis.
meatalinguistic knowledge (MLK), the relationship between LAAs and WM, as well as which
cognitive abilities relate to LAAs and whether or not LAAs are stable.

Measures of LAAs are strong predictors of L2 achievement. Li’s (2014) meta-analysis found
that measures of language analytic abilities show a small-medium correlation with L2
proficiency measures, \( r = .35 \) and 95% CI of \([.27–.43]\). Moreover, the results show that of all
the individual L2 aptitude components, language analytic abilities are the strongest predictor
of L2 achievement. However, the predictive ability of L2 aptitude measures is contingent
upon having appropriately matched criterion variables (Carroll, 1962). Therefore, LAAs will
not always be a useful predictor, given their dependence on variables such as instructional
treatment.

Research into learning variables and LAAs provides insight into the learning conditions
conducive to LAAs and the types of L2 learning which may result. LAAs are seen as a
potentially important variable in the efficacy of feedback. LAAs may be more important
when feedback is less structured (Jiang & Xiao, 2014; Sachs, 2010; Stefanou & Révész,
2015; Yilmaz, 2013), although this might be negated by feedback that includes metalinguistic
information (Jiang & Xiao, 2014; Stefanou & Révész, 2015; Yilmaz, 2013). Interestingly,
though, LAAs also seem to be more implicated in the acquisition of explicit L2 knowledge as
opposed to implicit (Robinson, 2005; Wistner, 2014). LAAs also seem to be more related to
receptive skills as opposed to productive skills (Erlam, 2005; Horwitz, 1980), particularly in
relation to literacy skills (Detschelt, 1993). However, LAAs do not seem associated with L2
processing tasks (VanPatten & Smith, 2014). In summary, these insights tend to suggest that
LAAs are most important when L2 input is less structured and learners have to actively
construct the L2 system themselves (Skehan, 2002). Actively analysing L2 input, therefore,
results in explicit knowledge of the L2 system, enabled by reading and writing skills related to tasks in which communicative pressures are lessened. However, the benefits of LAAs in L2 learning this way might be nullified by L2 instruction. Taking these findings into account, the training course in the current study was designed to raise the conscious awareness of L2 learners of how they could analyse L2 input and extract information in a way that would allow them to construct the L2 system. Various strategies and techniques were also explained and practice tasks given that allowed L2 learners to practice analysing L2 input and, as a by-product, develop explicit knowledge of the target L2 system.

Both language analytic abilities and metalinguistic knowledge have been variables of interest in SLA studies on the effectiveness of feedback. However, the findings of these studies have been mixed, with no clear relationship between the two constructs evident. Although measures of LAAs and MLK have been found to correlate, principal components analyses have loaded them onto different factors (Alderson et al., 1997; Roehr & Gánem-Gutiérrez, 2009; Wistner, 2014). Findings from L2 learning studies investigating feedback treatments with and without metalinguistic information have also been mixed. Measures of LAAs and MLK both predicted L2 achievement, but measures of LAAs only predicted L2 learning if feedback did not contain metalinguistic information (Jiang & Xiao, 2014; Sheen, 2007; Stefanou & Révész, 2015). Even in predicting L2 learning, the results are unclear, showing that combined measures of the two constructs can be strong predictors (Roehr, 2007), but also with MLK predicting procedural L2 knowledge and LAAs predicting receptive metalinguistic knowledge (Wistner, 2014). Roehr (2007) suggests that the connection between the two constructs is developmental, with language analysis during L2 learning leading to increased metalinguistic knowledge. In short, the relationship between LAAs and MLK is obviously complex and requires more research to understand how they relate to each other and to L2
learning. In the current study, metalinguistic knowledge was incorporated by asking questions about L2 grammar as a concept. For example, elements of explicit L2 knowledge raised in the training course focussed on the frequency of forms and the nature of relationships that defined a grammatical concept, as well as the similarities/differences with analogous concepts in English (see sections 4.4.2- 4.4.3 for details).

LAAs also appear to be related to other abilities implicated in L2 learning, such as working memory, general cognition, and L1 skills. Measures of LAAs and WM correlate (Howes, 1999; Yalçın et al., 2016), although they load onto different factors (Yalçın et al., 2016). This suggests that the two constructs may overlap, but are also largely independent. Measures of LAAs also correlate with measures of general intelligence (Feenstra, 1967; Wesche et al., 1982) and are related to other general cognitive abilities such as noticing (Ahn, 2012), set extinction – the ability to switch between (autonomous) linguistic systems (Cummins & Gulutsan, 1975) –, and even field dependence (Detschelt, 1993). As noted above, measures of LAAs also seem to be related to L1 skills such as literacy (Detschelt, 1993; Sparks et al., 2011). More recently, it has even been proposed that measures of LAAs like the LLAMA F measure explicit L2 learning ability (Granena, 2013; Yalçın et al., 2016). In short, LAAs seem to overlap yet be distinct from other constructs, such as WM and general intelligence, as well as being implicated in and a result of L1 skills.

In summary, the research investigating language analytic abilities offers considerable and useful insights. Compared with other components of L2 aptitude, measure of LAAs are in general the strongest predictor of L2 achievement. However, LAAs are most important in L2 learning when L2 input is less structured and instruction less explicit, although this was not necessarily the case in the current study. In particular, LAAs seem to be mediated by the
amount of metalinguistic information available in feedback. The construct of LAAs also appears to be complex, overlapping with WM and general intelligence and implicated in L1 skills, particularly those of literacy. The current study attempts to incorporate these findings by identifying specific areas where LAAs and general intelligence might overlap (see section 3.2.3), and by sampling university students with presumed high levels of L1 literacy (see section 4.2) in an attempt to control for this variable.

### 3.2.2.3 Summary

Language analytic abilities are poorly defined and explained in theoretical accounts of the construct. This is surprising, given the importance of measures of LAAs in predicting L2 achievement. However, the relatively superficial descriptions of grammatical sensitivity and inductive language learning provided by these theoretical accounts are challenged in practice by the range of tests of LAAs that also seem to measure these same constructs. Nevertheless, research into LAAs suggests that the construct is more complex than simple measures of grammatical sensitivity and inductive language learning. LAAs seem to relate to WM, general cognitive abilities, and even L1 skills in ways not yet fully understood. Given the state of LAA research, efforts to identify the various sub-components of LAAs and their respective roles might usefully advance our understanding of the construct.

### 3.2.3 A novel conceptualisation of language analytic abilities

As argued above, the construct of language analytic abilities lacks a clear description or explanation. Like L2 aptitude, LAAs are what their respective tests measure. However, the findings from research using measures of LAAs offer a more complex picture than the
constructs of LAAs tests propose. It seems worthwhile, therefore, to reconsider the construct of LAAs. In the following section, the theory and research are examined in more detail and the argument is made that language analytic abilities are a composite of various general cognitive abilities as applied to the task of L2 learning. This prospective model is then used to design and create the training intervention to investigate the trainability of LAAs (see 3.4.3).

The first and strongest claim that can be made is that LAAs are related to general cognitive abilities (Skehan, 2002). Empirical evidence has uncovered weak to strong relationships between measures of LAAs and general ‘intelligence’ from various statistical approaches (Carroll, 1962; Gardner & Lambert, 1965; Pimsleur, Mosberg, & Morrison, 1962; Sasaki, 1996; Wesche et al., 1982). Further evidence that LAAs, as part of L2 aptitude, can be described and measured from a general cognitive framework comes from two more recent L2 aptitude measures: the CANAL-FT (Grigorenko et al., 2000) and the Hi-LAB (Linck et al., 2013). Both of these tests have clear theoretical constructs based upon a general cognitive perspective of L2 aptitude. Their success as valid measures of L2 achievement lends support to the view that general cognitive abilities underpin LAA. Adopting this general cognitive perspective, previous studies and reviews of L2 aptitude can also be reinterpreted for clues as to the specific processes and mechanisms of LAAs.

The second (and least controversial) claim that can be made about LAAs is that they comprise a bottom-up and data-driven process of analysis (Skehan, 2002). Carroll (1973:5-6) describes inductive language learning, one of the dimensions of LAA, as “the ability to examine language material (in either auditory or printed form) and from this to notice and identify patterns of correspondences and relationships involving either meaning or
grammatical form”. Two things are clearly described here. First, Carroll views LAAs as processes whereby knowledge of L2 is inferred from input, so that while previous knowledge may be beneficial at any particular point in learning, it is not a prerequisite (Carroll, 1962, 1981). In other words, learners start with the input and construct L2 knowledge from their inferences. Second, Carroll specifically names noticing as a mechanism of LAAs. As a bottom-up process of analysis, it seems logical that noticing would be part of the LAAs construct, as would be related abilities, such as awareness and recognition. Indeed, these abilities related to noticing are also identified by Carroll elsewhere (Carroll, 1973).

Other general cognitive abilities are also identified in the literature on LAA. Carroll (1990:19) describes the MLAT IV as “a grammatical analogies test in which, for each item, the examinee is asked to find a grammatical construction, in one or more English sentences, that has a function analogous to that of an indicated word or phrase in a key sentence”. In other words, analogical reasoning is an ability implicated in LAA.

The third and weakest claim able to be made is that LAAs result in L2 knowledge in the form of concepts. This claim comes primarily from Carroll (1981:107-108), who suggests that “grammatical sensitivity is a basic aptitude or capacity to acquire grammatical concepts”. This idea of L2 knowledge as being conceptual in nature also follows from the argument that LAAs are a subset of general cognitive abilities. If we are to take a generalist view of L2 acquisition, then we might also accept that L2 knowledge would be in the same form as all other knowledge, concepts (see 3.2.3 for a more detailed discussion).

Viewing L2 knowledge representation as conceptual has significant implications. First, it allows us to hypothesise which abilities are likely to be involved in the creation of such knowledge, e.g. categorization, schematization, analogy, and distributional analysis.
(Tomasello, 2003), as well as the form in which that knowledge exists, e.g. exemplars, relationships, and explicit knowledge (Murphy, 2004). Second, it allows us to hypothesise the features of the input that lead to such knowledge, e.g. meaningful exemplars that can be analyzed (Ellis, 2012), as well as variation in such exemplars that allows for abstraction (Tomasello, 2003).

In line with the proposals discussed above, the current study proposes that LAAs be reconceptualised as a composite of general cognitive abilities that are applied to the task of L2 learning. On this construal, LAAs are primarily bottom-up processes involving a range of abilities such as noticing, awareness, recognition, and analogical reasoning. In other words, language analysis involves not just one ability, but multiple abilities deployed to acquire grammatical concepts. While it is possible that the number of potential language analytic abilities is quite large, it seems probable that a handful are more important (see Carroll, 1962; Pimsleur, 1966). In the next section, the LAAs which are most likely to be important in L2 learning are listed and defined.

### 3.2.4 List of potential language analytic abilities

As the brief review above makes plain, many possible abilities could be said to underpin language analytic abilities. However, by concentrating on the analyses of existing measures of LAAs, the current study assembled a list of abilities identified in the literature as being of primary importance in analysing L2 input for the purposes of learning the patterns and structure of a second language. The list is neither exhaustive nor clear cut – some abilities, such as problem solving, are more general and involve the use of other abilities in their application to analysis of L2 input. For each hypothesised ability of language analysis below,
a formal definition as well as a short description of its role in language analysis are provided (see section 4.3.3 for the application of abilities in language analysis tasks).\footnote{While the terms and definitions given here are as detailed as possible, they are not exhaustive, nor do they offer clear limits as to where one ability stops and another begins. Consequently, the terms and their explanations should be regarded as suggestive rather than definitive.}

**Recognition** can be defined as a feeling of familiarity that something has been seen before (Granena, 2013). In the case of language analysis, it can be hypothesised that the repetition of instances must first be identified as such for further analysis to occur. In other words, recognizing words and phrases (at the surface level) and grammatical structures (at the abstract level) across multiple instances in the input can trigger noticing.

**Noticing** can be defined as the “conscious registration of the occurrence of some event” with regard to “surface level phenomena and item learning” (Schmidt 1995:28). The noticing hypothesis claims that for input to become intake, it must be noticed. Hence, for patterns and structures to be analyzed, they must first come to conscious attention.

**Reasoning** is understood as “the mental process of deriving consequences from given information” (Johnson-Laird 2000:75). Patterns and structures that have been noticed will increase what is able to be known about the L2 system. For categorization and concept development, reasoning about input allows for the system to develop in terms of breadth and depth. While many types of reasoning are important (e.g. propositional, syllogistic, everyday reasoning (Galotti, 2013)), the three types that appear to be most useful in LAAs are analogical, inductive, and deductive.

**Analogical reasoning** is defined as “the process of identifying goal-relevant similarities between a familiar source analog and a novel, less understood target, and then using the correspondences between the two analogs to generate plausible inferences about the latter”
(Holyoak & Richland, 2015:223). To learn how to use a second language means to know how to use a first. A lot of L2 knowledge can be inferred by using L1 patterns and structures as analogies, particularly in the early stages of L2 learning. Consequently, a more detailed knowledge of the L1 will enable more effective analogical reasoning in relation to the L2. Furthermore, L2 knowledge can also be a resource for L3 learning (Jessner, 2008), and L2 plus L3 knowledge for L4 learning, and so on.

**Induction** is defined as “an inference from the particular to the general” or “inferences from the particular to the particular” (Johnson-Laird, 2012:145). Learning the L2 bottom-up from the input will involve inferring from one instance of the L2 to another, e.g. “I am from Australia” could be inferred from hearing someone else say “I am from Spain”. An inference may also extrapolate from the particular to the general, e.g. from the previous example the copula construction may be inferred.

**Deduction** is understood as the reaching of a conclusion for a specific scenario based on a premise in general (Johnson-Laird, 2000). As the L2 knowledge system develops, information already known about the L2 may be used to process input from the top down. Such information may originate from previous inferences or from metalinguistic knowledge acquired consciously, e.g. in class or from a textbook.

**Problem solving** is construed as the overcoming of obstacles to answer a question or achieve a goal (Reed, 2000). On this understanding, problem solving is a general ability that can deploy the abilities already described above. Whether the goal is to communicate with speakers of the L2 or to acquire declarative knowledge of the L2 system as part of a formal language course, L2 learners need to overcome many obstacles. With such goal-oriented learning, solving problems of communication and/or answering questions about the L2
system are likely to occur and to offer L2 learners the opportunity to reinforce existing L2 knowledge through recognition, noticing, and reasoning. In addition, successful problem solving in communicative situations depends on metacognitive knowledge of the topic that allows for the effective and timely deployment of strategies (Reed 2000, see 3.2.2.2 for discussion of metacognitive knowledge and strategies in L2 aptitude). Problem-solving abilities are also expected to vary substantially in both number and degree between individuals, rendering problematic their delimitation as a simple process.

Problem solving abilities are hypothesised to operate whenever the learner analyses L2 input. However, it must be noted that these abilities are differentially implicated at different stages of acquiring L2 knowledge. Therefore, the current study proposed a model of LAAs used in acquiring L2 grammatical concepts (see figure 3.4), that accords with the generalist view of L2 knowledge described above (see 3.1.3). This model, while still to be substantiated, provides the theoretical foundation for designing the training course in the current study (see sections 4.4.2-4.4.6).
Figure 3.4 Speculative model of language analytic abilities in acquiring L2 knowledge

The following is a description of the model in figure 3.4. Starting from the top-left, from the input the learner becomes aware of a certain grammatical pattern or patterns (GP). The learner then turns their attention to this pattern in the input and begins to notice instances of the GP in other input. From these instances, the learner begins to infer a concept of the GP, although the concept is probably fuzzy and not too defined at this point. Based on this concept, the learner develops certain expectations of how the GP will present in the L2 and these hypotheses are activated when the learner attends to more input. The new input will present instances of the GP that the learners can use to confirm or challenge their hypotheses. From this new input, the concept of the GP will clarify and become more concrete for the learner. From here it will influence the L2 competency and/or L2 proficiency of the learner.
This in turn will feedback into the input of the learner and again meet or challenge expectations of the GP and continue to clarify or restructure, if necessary.

While this speculative model informs the design of the training tasks used in the current study, its accuracy is beyond the scope of the present thesis. For our present purposes, suffice it to note that the model provides a theoretical framework of processes and mechanisms of language analysis used to design and create the training tasks of the current study.

The current study views the acquisition of a L2 grammatical system as involving constructing and utilizing grammatical concepts. Language analytic abilities underlie and drive this acquisitional process. Put otherwise, LAAs are general cognitive abilities applied to the domain-specific task of processing/analysing L2 input to structure it as part of building and internalising the L2 system.

At their most basic level, LAAs can be summarised as general problem-solving abilities which include (1) attending to; (2) noticing patterns in; and (3) thinking deeply about similarities and linguistic relationships present in the language input. The current study proposes that these three abilities are central to language analysis, which in turn drives the development of L2 grammatical knowledge and skills.

The current study proposes that language analysis comprises five interdependent stages in a feedback system. These five sequential stages are: awareness, concept formation, hypothesis formation, hypothesis testing, and concept clarification/restructuring. The training course of the current study is structured around these five stages, and is informed by the hypothesis that these five stages require the learner to strategically apply all language analytic abilities in response to a L2 learning task/problem.
Having established the theoretical underpinnings of the current study’s training course, we will turn in the following section to its specific tasks (see 4.4.3), content (see 4.4.4-4.4.5) and overall structure (see 4.4.6).

3.2.5 Summary of language analytic abilities and the implications for training

Evidently, our discussion has highlighted that current theories of language analytic abilities are firmly based on current measures of the construct and lack definitional clarity. Put otherwise, LAAs are what LAAs tests measure. At the same time, tests of LAAs vary in form and task structure, despite purporting to measure the same construct. The research confirms that LAAs are more complex than the simplified descriptions of early tests, such as the MLAT IV Words in sentences and the PLAB 4 Language analysis sub-tests. LAAs appear to overlap with several constructs, yet are also independent of them at the same time. Given this definitional confusion and evident complexity, new perspectives on LAAs are needed to offer explanations that are more detailed and that better accord with the empirical findings to date.

The current study proposes that language analytic abilities be reconstrued as a composite of general cognitive abilities applied to the specific task of L2 learning. These cognitive abilities are potentially large in number, but a subset that facilitates bottom-up processing of L2 input for the purposes of extracting grammatical patterns and structures has been proposed. The result of this bottom-up analysis of L2 input are grammatical concepts that we use to structure and interact with our L2 environment. Consequently, a novel model has been proposed that describes how language analytic abilities operate for the purposes of L2 learning.
To test the trainability of LAAs, the current study adopted the model above, informed by both theory (see section 3.2.1) and research (see section 3.2.2), in order to develop a training course, details of which are provided in sections 4.4.2-4.4.6.

3.3 Research questions

Previous L2 aptitude research has given rise to suggestive but inconclusive findings with regards to the assumptions of stability and untrainability. In some pre-/post-test studies, findings of gains in L2 aptitude scores (Ganschow & Sparks, 1995; Gliksman et al., 1979; Sáfár & Kormos, 2008; Sparks & Ganschow, 1993; Sparks, Ganschow, Artzer, & Patton, 1997; Sparks, Artzer, Patton, Ganschow, Miller, Hordubay, & Walsh, 1998; Sparks, Ganschow, Fluharty, & Little, 1995; Sparks, Ganschow, Pohlman, Skinner, & Artzer, 1992) are questionable due to maturational effects (Gliksman et al., 1979). Other pre/post-test studies have lacked statistical evidence to substantiate claims of gains, with a lack of paired-sample comparisons (Politzer & Weiss, 1969; Sáfár & Kormos, 2008; Yeni-Komshian, 1965) or statistical tests of any kind (Demuth & Smith, 1987). In addition, these studies did not report effect sizes or include any explanatory or predictive modelling of gains, instead favouring group comparisons. Such study designs provide no insights into the relative importance of the variables responsible for gains in L2 aptitude scores. Lastly, the two previous L2 aptitude training studies (Politzer & Weiss, 1969; Yeni-Komshian, 1965) used a primarily deductive approach to learning in which much of the content was based on (meta)linguistic knowledge.

Consequently, the following research question motivated the current study:
1. Is language learning aptitude stable or is it sensitive to experience?

To explore this question, a training program was developed, focusing on whether LAAs, a crucial aspect of L2 aptitude, could be enhanced through training. LAAs were then tracked in relation to phonological and memory components of L2 aptitude. On this basis, the following two secondary questions allow the primary research question to be answered:

a. **Using a pre-/post-test design, do post-test scores for L2 aptitude differ significantly from pre-test scores?**

b. **Do training variables explain more variance in post-test scores for language analytic abilities than L2 aptitude variables?**

Building on the findings of previous studies, the current study seeks to deepen our understanding of these issues. First, to minimise any maturational effects, participants are sampled from an adult population of university students, not from a teenage population of high school students. As previous studies have in general investigated university-aged participants (see section 2.2.4.4), sampling from comparable populations increases the generalizability of the findings from the current study with the rest of the field. Second, the study uses paired-sample *t*-tests to measure gains within subjects and quantify any gains by way of effect sizes, calculated with Cohen’s *d*. Third, to offer insight into what may contribute to any potential gains in test scores of language analytic abilities, predictive models will be constructed by collecting data on a range of variables (see section 5.2). Finally, to account for confounding issues identified in Politzer & Weiss’s (1969) previous L2 aptitude training study (see section 3.1.2.3), the training course in the current study was based on explicit theoretical assumptions and designed to be simpler, more focussed on a specific area of L2 aptitude abilities, integrated with but distinct from participants’ actual L2
(Spanish) learning, and targeting higher-order aptitudinal sequencing and strategies (see section 4.4.4 for details).
4 METHODOLOGY

The current study investigates the issues of stability and and its correlate, untrainability, in L2 aptitude by employing a pseudo-experimental field study design. The study investigates stability by analysing for possible changes in L2 aptitude between pre- and post-testing. Untrainability is investigated by designing and implementing a training course in language analytic abilities and subsequently analysing its efficacy. The current chapter details the methods and materials employed in carrying out these two lines of investigation.

This chapter begins with an overview of the research design (section 4.1) and a description of the participants (section 4.2). Section 4.3 details the procedure the current study followed. Section 4.4 describes the materials, including the L2 aptitude instruments (section 4.4.1), an overview of the language analytic abilities training (section 4.4.2), the training task types (section 4.4.3), instructional materials developed for the training (section 4.4.4), the linguistic structures covered in the training (section 4.4.5), and the syllabus of the training course (section 4.4.6). The final three sections cover data collection (section 4.5), an overview of the analysis employed in the study (section 4.6), and the limitations identified (section 4.7).

4.1 Research Design

The study was longitudinal, following a simple pre-test, treatment, post-test design. No control group was included, as creating a true control group was impractical for the following reasons. Time on task has been identified as an important variable in learning and in training aptitude (Jaeggi et al., 2008). To make treatment and control groups comparable in this respect, they would need to have spent comparable amounts of time on unrelated learning tasks. Plainly, this poses two problems. First, any learning task selected for the control group
would need to be both relevant to L2 learning and motivating, but also independent of L2 aptitude. Finding a learning task for the control group that would in no way impact on aptitude yet be relevant to L2 learning was extremely difficult. Second, a more important consideration was that participants in the study self-selected how much of the training they would complete and, therefore, how much time they spent on task. Allowing participants to make choices about training themselves was intended to maximise the chances that attitudinal and motivational factors would not confound performance on the training course. The rationale here was that if participants were free to control their own training, then they would train when they were motivated to do so, and higher motivation would lead to better performance over the course of the training. However, creating matched control and treatment groups with this self-selection was a methodological problem beyond the capacity of this study. In addition, a rigorous training schedule imposed externally on learners would not reflect learning in the real world, would have decreased the environmental validity of the study, and potentially increased the chances of motivational issues confounding the data (Politzer & Weiss, 1969). There were also ethical considerations to take into account (see section 4.2.1). For these reasons, no control group was included in this study.

With the design of the study now explained, we move onto the description of participants in the study.

### 4.2 Participants

The participants in this study were university students taking a first-year introductory Spanish course in a metropolitan area on the Australian east coast. Of the 128 students who initially agreed to take part in the study, 115 completed all four of the LLAMA pre-tests. Of the 115
students who completed all the LLAMA pretests, 90 completed post-tests for the LLAMA B, D, and F, and 85 completed all four of the L2 post-tests. All statistical analyses carried out utilised the maximum sample size possible and are detailed for each test.

Table 4.1 shows the age of participants. The majority of participants were young adults between the ages of 20 and 25 years old. This age group was sampled to minimise any possible maturational effects on L2 aptitude testing (see section 3.1.2.3). Ages ranged from 17 to 63 years old, with an average of 23.87 (s.d. = 8.22 years), and an interquartile range of 20-25 years old. Six participants did not report their age. Gender data were not collected.

Table 4.1. Descriptive statistics for age of participants

<table>
<thead>
<tr>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Range</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>115</td>
<td>23.87</td>
<td>8.22</td>
<td>21</td>
<td>17-63</td>
<td>20-25</td>
</tr>
</tbody>
</table>

Table 4.2 shows participants’ first language and the number of speakers. Participants were predominantly L1 English speakers (77%), although a range of L1 backgrounds was reported. The remaining 23% of participants reported a wide variety in first languages, both typologically and geographically.

Participants also reported a range of previous experience with learning other languages, reported in table 4.3. A measure of language learning experience was collected by asking participants to report if they spoke any additional languages and, if so, to name those languages. Many participants self-reported proficiency levels as well as the languages, so a level of language experience was created based on the findings of the studies discussed in
This measure of L2 learning experience was an attempt to capture a measure of knowledge of L2 learning (see section 3.1.2.2).

Table 4.2. Participants’ first language backgrounds

<table>
<thead>
<tr>
<th>n</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>English</td>
</tr>
<tr>
<td>4</td>
<td>Japanese, Norwegian</td>
</tr>
<tr>
<td>2</td>
<td>Arabic, German, Swedish</td>
</tr>
<tr>
<td>1</td>
<td>Afrikaans, Bahasa Malay, Bengali, Cantonese, Dutch, French, Greek, Italian, Kannada, Kiribati, Mandingo, Mandarin, Motu, Polish, Russian, Tagalog, Uyghur</td>
</tr>
</tbody>
</table>

Over half the participants (53%) self-reported as being language learners with no previous experience learning another language; approximately a third (34%) self-reported as having started or learned a second language; and just over a tenth (12%) self-reported as having learned more than a single second language.
Table 4.3 L2 experience levels of participants

<table>
<thead>
<tr>
<th>L2 experience level</th>
<th>Number of participants</th>
<th>Self-reported L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>61</td>
<td>Beginner Finnish, Beginner French, Beginner Japanese, Beginner Sanskrit, Beginner Spanish</td>
</tr>
<tr>
<td>Low</td>
<td>8</td>
<td>Arabic, Cantonese, English, French, Japanese, Lingala, Mandarin, Polish, Portuguese, Tuvaluan</td>
</tr>
<tr>
<td>Some</td>
<td>32</td>
<td>Bahasa Malay, Cantonese, Danish, English, Flemish, French, German, Hindi Marathi, Hokkien, Mandarin</td>
</tr>
<tr>
<td>Extensive</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

Participants also reported varying levels of university education. The vast majority (92%) of participants were taking undergraduate degrees. Of the five participants who reported undertaking a diploma\(^{16}\) in second languages at the university, one self-identified as undergraduate and another as postgraduate. One participant reported taking “Other” levels of study without any further specification. In general, the education level of the sample reflects well developed literacy skills, a variable that has correlated with measures of language analytic abilities (Detschelt, 1993; Sparks et al., 2011). Generally high levels of L1 literacy across participants suggest that this variable is less likely to confound results (see section 3.2.2.2 for discussion).

Participants reported on their professional (i.e., not casual or unskilled) work experience. These data were collected not only to characterise participants’ formal education, but also to indicate cognitive abilities potentially acquired from work experience, such as problem-solving.

---

\(^{16}\) A diploma here refers to a university course lower than a undergraduate or postgraduate qualification. A diploma can be a lower level qualification or can be extended to become part of an undergraduate or postgraduate degree with further study. The university in this study offered diplomas in second languages.
solving, planning, or what Sternberg (2002) terms “practical intelligence”. Just over half the participants (55%) reported having never worked professionally before, while only 7% reported having worked > 10 years, 6% < 5 years, and 5% for 5-10 years.

Participants’ reported motivations for learning Spanish were varied. In the background questionnaire, participants were asked to select one of five motivations for learning L2 Spanish, including: “To improve international job prospects”, “For travel”, “For fun”, “As an intellectual challenge”, and “To meet and make Spanish-speaking friends”. Participants were able to select as many of these motivations and were asked to list them in order of importance. The median number of motivations selected was three. The most frequent primary motivation was “For fun”, selected by 59% of participants. The most frequent secondary motivation was “For travel”, selected by 50% of participants. The most frequent tertiary motivation was “To meet and make Spanish-speaking friends”, selected by 34% of participants. Overall, the most common motivation for learning L2 Spanish, regardless of order of preference, was to “Improve international job prospects”, selected by 74% of participants. The other most common motivations, in order, were “For travel” (64%), “For fun” (59%), “As an intellectual challenge” (58%), and “To meet and make Spanish-speaking friends” (47%). Overall, participants’ motivations were multiple and mixed, including intrinsic and extrinsic motivations. Primary motivations tended to be intrinsic (“For fun”, “As an intellectual challenge”, “To meet and make Spanish-speaking friends”); secondary motivations tended to be extrinsic (“To improve international job prospects”, “For travel”); while overall totals for each type of motivation were fairly even.

All participants were undertaking an introductory Spanish course at the university. The method of instruction was primarily communicative and centred around the course textbook.
Dos Mundos 7th ed. (Terrell, Andrade, Egasse, & Muñoz, 2010). Participants attended two two-hour lessons each week for 13 weeks and in addition were given weekly homework that included an explicit grammatical knowledge component. Lessons were primarily conducted in Spanish, but some class teachers made allowances for English in the classroom to facilitate learning when deemed appropriate. Assessment for this course involved oral exams testing communicative and linguistic competence, as well as written exams testing listening and reading comprehension, writing skills, and grammatical knowledge.

4.2.1 Ethical considerations

The sample was one of convenience, to maximise participation in the pre-/post-test study. All students (100%) taking first-year introductory Spanish in semester 1 were asked to participate. Participation involved consenting to the research team accessing data from the Language Learning Skills & Strategies (LLSS) assessment items, which included all the measures used in the study. All students, regardless of participation in the study, were expected to complete the LLSS as part of their introductory Spanish course. The LLSS was worth 5% of a student’s grade and was a formative assessment item. This meant that if a student were to complete 100% of all tasks in the LLSS assessment items, they would receive 5% towards their final grade. If they completed 80% of the LLSS, a student would receive 4% towards their final grade. Participation in the study was completely voluntary and consenting for data to be used in the study had no impact on final university course grades. Participants volunteered to have their data included in the project, self-selecting themselves for the study. All data were then collected via the university online learning management
The project was conducted in accordance with Griffith University ethical guidelines and approval (Project Reference No. LAL/05/14HREC). Students undertook the Language Learning Skills and Strategies (LLSS) assessment online via the university learning management system. Upon beginning this assessment, participants were first presented with the informed consent procedure. Students were taken to a video that explained the research project, directed them to the informed consent document for details, and then asked for their consent to allow the research team to access their assessment data. The informed consent document was accessible via a clear link.

To indicate whether they had given consent or not, students needed to answer the consent question – “Do you understand and do you agree for the anonymised data from this assessment to be used for research purposes?” – by selecting “Yes” or “No”. Each student needed to answer this question before being able to access and begin their assessment item. If a student failed to clearly indicate their response, the system did not allow them to continue.

The principal researcher also manually went through all responses to ensure that all students had indicated whether they had given their consent or not. Of the 141 students enrolled in the introductory Spanish course, 128 (91.4%) voluntarily consented to the research team accessing data from the LLSS; six (4.3%) students declined to volunteer their data; and six (4.3%) students did not respond to the consent question.

As mentioned in section 4.1, one of the reasons for not having a control group was due to ethical considerations. In applying for ethical approval, the issue of equality regarding a control group was raised. Because data were harvested from a formative assessment item, all
students needed to have equal access to the same learning materials. This prohibited the inclusion of a control group, as the control group would need to be excluded from accessing any language analysis training materials.

4.3 Procedure

This section discusses the procedure participants were required to follow as part of the Language Learning Skills and Strategies (LLSS) formative assessment item. As mentioned, the LLSS was part of the first-year Spanish course. Although all students were required to complete this LLSS assessment, the data for the study were collected only from students who had volunteered.

Participants first completed a questionnaire asking for background information: age, education, professional experience, first language, and second language(s)\(^\text{17}\). They then completed the LLAMA pre-test of second language aptitude by accessing a download link for a program that packaged and executed the LLAMA test (Meara, 2005) (see section 4.4.1 for more details). Participants were explicitly instructed on how to run this program, complete each sub-test, and then enter and upload their results to the university’s learning management system (LMS). Additional information was provided to participants about the LLSS assessment item, including: assessment overview, assessment video explainer, and additional instructions on how to access and run the LLAMA test.

Students had until the end of the second week of the semester (week 2) to initially access, agree to participate in the study, complete the background questionnaire, and complete the

\(^{17}\) Note again that some participants also self-reported proficiency levels (see table 4.3).
LLAMA pre-test. In total, 13 of the 128 participants who had initially volunteered did not meet this deadline and were excluded from the study (final sample = 115).

In week 2 of the introductory Spanish course, participants were given access to the first week of LAA training materials in the LLSS assessment. The procedure for each week was the same, with the materials for the week being made available on the Monday morning. While participants could not access training materials for weeks to come, they could access training materials for all previous weeks up until the end of the data collection period in week 9 of the Spanish course.

Materials for each week consisted of a short video explainer and the training tasks. Each video explainer dealt with one aspect of LAAs and its related skills and strategies for L2 learning (see section 4.4.4 for more details). In addition, a range of two to five training tasks that focused on the relevant skills and strategies for that week (see section 4.4.3 for more details) was made available to participants. The materials for each week were designed to take approximately 30-35 minutes a week in total to complete. Participants were encouraged to do a little bit every day and to space their training, but were free to choose how they engaged with the training materials. Most opted to do all the training for the week in one massed session once a week or once every few weeks.

Participants completed the training materials over the course of six weeks of their course, not including a week of university holidays. Although access to the training was not suspended during this time, no new materials were made available to participants and data showed little activity from participants during the holiday week. At the end of the six weeks of training, which equated to week 7 of their Introductory Spanish course, participants were sent the same materials for the LLAMA post-test as they had received for the pre-test. Once again,
they were given two weeks within which to complete this test and upload their results to the university’s LMS. At the end of this 9-week period, online access to the LLSS materials used for data collection was closed.

Once the online data collection had been completed, all raw data were downloaded for collation and analysis. Data were then collated into a master spreadsheet that contained all the measures of all pertinent variables (see section 5.2 for more details). Data analysis was carried out in the R programming language (R Core Team, 2016) with the assistance of an experienced statistician. Paired-sample $t$-tests were employed to answer research question 1.a. regarding stability in L2 aptitude and a random forest regression model was employed to answer research question 1.b. regarding the efficacy of targeted training (see section 4.6 for more details).

4.4 Materials

The following section describes the measures and training materials used as part of the study. Section 4.4.1 describes the LLAMA L2 aptitude tests and their administration in the study. Section 4.4.2 describes the language analytic abilities training and its materials, including the training task types (section 4.4.3), the instructional video explainers (section 4.4.4), and the linguistic structures of Spanish included (section 4.4.5).

4.4.1 L2 aptitude test instrument

L2 aptitude was operationalized as scores on the LLAMA aptitude test (Meara, 2005) (hereafter referred to as ‘the LLAMA’). Meara (2005) states that the LLAMA is exploratory
in nature and not suitable in high-stakes situations. However, the LLAMA was deemed suitable for use in the current study due to (a) its previous use in L2 aptitude research (see Artieda & Muñoz, 2016; Granena, 2014; Granena & Long, 2012; Suzuki & DeKeyser, 2016; Yalçın et al., 2016; Yalçın & Spada, 2016; Yilmaz, 2013) and (b) its acceptable levels of reliability – internal consistency $\alpha = .66–.77$ and test-retest reliability $r = .53–.64$, $p < .05$ over two years (Granena, 2013:122) (although see section 3.1.2 for commentary on the strength of these levels of test-retest reliability). The digital format of the LLAMA, enabling its administration over the Internet, also avoided the difficulties of requiring all participants to come to a central location at an organised time for pre- and post-testing. Although not without its problems (see sections 5.2.1.1-5.2.1.4 for further discussion), this was an attractive option for maximising the possible sample size.

The LLAMA test is a suite of four computerised tests available for free download at the lognostics website (http://www.lognostics.co.uk/tools/llama/index.htm), which is an online resource for research tools managed by Professor Paul Meara. The four tests, described in more detail below, are LLAMA B (vocabulary learning), LLAMA D (sound recognition), LLAMA E (sound-symbol association), and LLAMA F (grammatical inferencing). In total, the four LLAMA tests take approximately 30-60 minutes to administer and are designed to run on Windows XP operating system and above. The LLAMA test programs automatically mark each individual test and convert raw scores to a standardised score on a scale of 0 or 5 to 100, in increments of 5 for most tests. Administration of the tests is also relatively standardized, as the test runs automatically once certain parameters have been set. To maintain the security of this standardised administration and ensure that participants did not interfere with these parameters, the LLAMA test programs were bundled into one file that
executed automatically one after the other with the parameters already set according to the
instructions in the LLAMA manual (Meara, 2005).

To bundle the four separate LLAMA test programs, a wrapper was created around the
LLAMA software based on functionality provided by other software, 7zsfx
(http://7zsfx.info/en/). This software turns simple file archives into self-extracting ones and
adds some functionality to the extraction process. The process to wrap the LLAMA software
was done in three steps:

1. The LLAMA executables were re-packaged into a single archive in the 7z format (see
   http://7-zip.org/).
2. A configuration file for 7zsfx was created.
3. The 7zsfx software, the configuration file and the LLAMA archive were combined by
   concatenating them into a stand-alone executable (named “ProfileTest.exe”).

The 7sfx software was configured to:

1. extract the contained archive (with the LLAMA software) to a temporary location on
   the user’s computer;
2. show an initial prompt to the users;
3. run the four LLAMA programs one after the other from where the archive was
   extracted to;
4. run another program to copy the LLAMA result file to the user’s desktop (with a
different extension so they would not realise it contained only text); and
5. clean up, i.e. remove the extracted files and finally remove itself.
This left participants with just the LLAMA results file on their desktop, which they were requested to report online before attaching and sending the file to the researcher via the university LMS. This final step also allowed for the researcher to double check self-reported test scores for inconsistencies while collating the data.

In the LLAMA manual, Meara (2005) describes each sub-test and its development. In general, the LLAMA is loosely based on the Modern Languages Aptitude Test (Carroll & Sapon, 1959). The tests are language independent (Granena, 2013) and utilise picture stimuli with verbal material adapted from North- and Central-American languages that most test-takers are unlikely to be familiar with. Like the MLAT, the LLAMA is a series of work-sample tests where test-takers are required to learn an aspect of an unknown language. Each test follows the same essential structure of an initial learning period followed by a testing period and feedback for answers given is in the form of a ‘ding’ sound for a correct answer and a ‘bleep’ for an incorrect answer. Each test is described in more detail below.

**LLAMA B** is a test of vocabulary learning, loosely based on the MLAT-V Paired Associates test (Carroll & Sapon, 1959), which measures the ability to learn new vocabulary in a short amount of time. In the learning stage, test-takers are presented with a matrix containing 20 images that represent 20 real words from a Central-American language (see figure 4.1). Note taking is not allowed, although the current study was unable to control for this. When test-takers click on an image, the name for that image is presented in a rectangle in the centre of the screen. Test-takers have two minutes to learn all the word-image associations. In the testing stage, in the rectangle in the centre of the screen test-takers are presented with the same 20 names from the learning stage but in a random order. Test-takers identify and select
(by clicking on) the correct image for each name. For each correct answer, 5 points are awarded, with no deductions for incorrect answers.

**Figure 4.1 LLAMA B test console**

**LLAMA D** is a test of sound recognition, loosely based on the work by Service (Service, 1992; Service & Kohonen, 1995) and Speciale, Ellis, and Bywater (2004), measuring the ability to recognise patterns in speech. In the learning phase, test-takers hear 10 short computer-generated sound sequences in an unknown language. Immediately after the learning phase, test-takers hear another 10 sound clips, some of which are repeated from the learning phase and others which are novel. For each sound clip, test-takers indicate if it contains one

---

18 “Napoleon Bonaparte” is the default name entered for the test-taker in the LLAMA aptitude tests.
of the 10 sounds from the learning phase (see figure 4.2). For each correct answer, 5 points are awarded and 5 points deducted for each incorrect answer.

**Figure 4.2 LLAMA D test console**

**LLAMA E** is a test of sound-symbol association, similar to the MLAT-II Phonetic Scripts test (Carroll & Sapon, 1959), which measures the ability to associate sounds with written symbols. In the learning phase, test-takers are presented with a panel containing 24 symbols (see figure 4.3). Each symbol is a button, which when pressed plays a sound. Test-takers have two minutes to learn the sound-symbol pairings and were instructed that they may take notes to help them do so, as per the test manual. In the testing phase, test-takers hear a word consisting of two syllables from the learning phase and are then presented with two possible spellings for the spoken word. Test-takers click on the spelling that matches the word they
heard. For each correct answer, 5 points are awarded and 5 points deducted for each incorrect answer.

![LLAMA E test console](image)

*Figure 4.3 LLAMA E test console*

**LLAMA F** is a test of grammatical inferencing, vaguely similar to the PLAB 4 Language Analysis test (Pimsleur, 1966). The LLAMA F test measures the ability to inductively learn the rules of an unknown language. In the learning phase, test-takers are presented with a panel of buttons, which when clicked display a picture and a phrase that encodes the relationship shown in the picture (see figure 4.4). Test-takers have 5 minutes to learn the rules of the language and were instructed that they may take notes to help them do so, as per the test manual. Test-takers click on the sentence they believe to be grammatically correct. For each correct answer, 5 points are awarded and 5 points deducted for each incorrect answer.
The LLAMA tests were used as the instruments for pre- and post-testing. LLAMA Total was also operationalised in the current study simply as the sum of the LLAMA B, LLAMA D, LLAMA E, and LLAMA F test scores. This allowed for an aggregate measure of L2 aptitude beyond the abilities measured by the individual sub-tests (Snow & Lohman, 1984). Having described the LLAMA test instruments, we will describe and discuss the materials for the training phase of the study in the following section.

4.4.2 Language analytic abilities training

The current study focussed on LAAs and adopted a novel perspective of the construct based on the theoretical and empirical insights discussed above (see section 3.2.3). A training course was developed based on this model of language analysis. A key part of this model is
that LAAs is itself an umbrella term for a range of abilities that may be used in the analysis of language for the purposes of second language learning. On this understanding, LAAs constitute a componential, rather than a unitary, construct. One of the goals for the training course, therefore, was to present a range of tasks that involved various types of language analytic abilities.

The current study’s LAA training consisted of six weeks of explicit instruction and training, all of which was conducted online. The training course included a range of general cognitive abilities hypothetically related to language analysis for L2 learning (see sections 3.2.3 and 3.2.4). Each week, participants were expected to spend approximately 30-35 minutes learning about how to analyse language for L2 learning and applying this knowledge in a range of short training tasks analysing Spanish grammatical concepts problematic for L2 learners. Training course content was extracted from participants’ Spanish course textbook and the grammatical concepts were introduced concurrently with their in-class treatment, as much as possible.

Following the list of possible LAAs identified in chapter 3.2.4, a variety of tasks was designed and developed for the purposes of training. These training tasks are described in the following section.

4.4.3 The training tasks

In total, eight different tasks were designed and created for the purposes of training the study participants in LAAs. These eight tasks incorporated each of the presumed abilities of language analysis: recognition, noticing, analogical reasoning, inductive reasoning, deductive reasoning, and problem solving (see section 3.2.4 for definitions and explanations).
variety of tasks was also intended to keep participants more engaged over the course of the training. The training tasks were designed to become more complex and difficult as participants presumably progressed in the LAAs and their knowledge of Spanish improved. The increase in complexity and difficulty was incorporated by adding longer phrases and sentences to analyse for the target grammar, as well as by including target grammar in each week which was more difficult than that included in the previous week (see section 4.4.5 for the target grammar and table 4.6 for an overview of the training course syllabus).

Below follows a description of each task, what motivated the design of each task, where each task was expected to be most important in the model of language analysis (see section 3.2.4 for details and figure 4.5 below – copied from section 3.2.4), and which abilities were most implicated.
4.4.3.1 Pattern Matching

Figure 4.6 shows an example of the pattern matching training task. The participant is first presented with a grammatical pattern in both English and L2 (Spanish). The grammatical pattern is presented by highlighting exemplars of the pattern in the sentences. The participant’s task is to identify and learn the pattern. Participants are then presented with novel sentences in the L2 in which the same pattern has to be located. The words that follow this pattern then need to be written. Translation equivalents in English are also presented in the question sentences to minimise lexical problems.

This task was inspired by both the MLAT IV (Carroll & Sapon, 1959) and the PLAB 4 (Pimsleur, 1966) and attempted to combine elements from both into a single task. The task was designed to promote noticing of grammatical patterns (Carroll, 1973) (verbal morphology in the example below) and analogical reasoning in moving from the exemplar sentences to those in the question (Carroll, 1990). Language analysis of grammatical patterns was explained to participants as part of the initial stages of the proposed model (see figure 4.4) where the learner becomes aware of and constructs (fuzzy) concepts.
4.4.3.2 Problem Solving

Figure 4.7 shows an example of the problem solving training task. The participant is given a small sample set of language in Spanish with translations into English. From this given material, the participant must write novel words or phrases by recombining the grammatical patterns from the input.

This task was inspired by the linguistic problem sets from the Australian Computational and Linguistics Olympiad (Estival et al., 2014) (past problem sets can be found at http://ozclo.org.au/past-problems/), which are not dissimilar from the PLAB 4 Language analysis test. The task was designed to promote the induction of grammatical patterns (Skehan, 2002) (plural morphology in the example); analogical reasoning (Carroll, 1990) across English and Spanish; reasoning about whether patterns evident in the given data are
generalizable to new contexts and therefore predictable (Pimsleur et al., 1962); and that the
data given pose a problem which can be identified, structured, and then solved (Reed, 2000).
Language analysis as problem solving was explained to participants as part of the hypothesis
testing stage of the model (see figure 4.5).

<table>
<thead>
<tr>
<th>Español</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Un hombre tiene dos brazos, dos ojos, dos pies y dos piernas.</td>
<td>A man has two arms, two eyes, two feet, and two legs.</td>
</tr>
<tr>
<td>Mi profesor de español sólo tiene un pie y una pierna.</td>
<td>My (male) Spanish teacher only has one foot and one leg.</td>
</tr>
<tr>
<td>La luz en la pared es azul.</td>
<td>The light on the wall is blue.</td>
</tr>
<tr>
<td>Escribo con un lápiz y un borrador.</td>
<td>I write with a pencil and an eraser.</td>
</tr>
<tr>
<td>Una mujer joven con ojos azules y pies pequeños es bonita.</td>
<td>A young woman with blue eyes and small feet is beautiful.</td>
</tr>
<tr>
<td>Un hombre con brazos largos y orejas grandes es feo.</td>
<td>A man with long arms and big ears is ugly.</td>
</tr>
<tr>
<td>Las mujeres viejas son pesimistas.</td>
<td>Old women are pessimistic.</td>
</tr>
<tr>
<td>Los profesores interesantes escriben con lápices en las paredes.</td>
<td>Interesting teachers write with pencils on the walls.</td>
</tr>
</tbody>
</table>

**Figure 4.7 Problem solving training task example**

### 4.4.3.3 Grammaticality Judgement Task

Figure 4.8 shows an example of the grammaticality judgement training task (GJT). The participant is presented with a sentence in Spanish and must decide whether the sentence is
grammatically acceptable (Yes) or unacceptable (No). All instances of this task have a time limit, except for the first time that the participants were presented with this type of task. Adding the time pressure after initial task exposure was intended to increase the difficulty level.

This task was introduced as an opportunity for L2 learners to test their L2 knowledge, as it is a common measure of L2 knowledge and L1 competence (R. Ellis, 2004). It was designed to promote the recognition of previously seen L2 constructions (Granena, 2013) (many of the lexical items in the GJT were recycled from prior training tasks) and the noticing of grammatical patterns and constructions (Carroll, 1990) salient to the grammatical concept being learned in the training for that week. Language analysis that involved grammaticality judgements was explained to participants in the training as part of both the concept formation and hypothesis testing stages of the model (see figure 4.5).

**QUESTION 1**

*Ashley y Mary Kate Olsen son ricos y bonitas.*

- [ ] Yes
- [ ] No

*Figure 4.8 Grammaticality judgement training task example*

4.4.3.4 *Multiple Choice*

Figure 4.9 shows an example of the multiple choice training task. The participant is presented with an incomplete sentence in Spanish that must be completed by selecting the most appropriate of four possible answers.
While this is a relatively straightforward task, the questions were designed to stimulate analysis of very specific elements of the target grammatical concept (see 4.4.5 for a list and details). The task was designed to promote the problem solving skills of identifying, structuring, and then solving (Reed, 2000) as well as reasoning between the alternative solutions. As the possible answers themselves offer clues as to the nature of the problem, a type of strategic thinking was also expected to be involved in which participants reflect on how the answers might offer insight into the task at hand. Language analysis as this type of reflective reasoning was explained to participants as part of the reconstruction/clarification stage of the model (see figure 4.5).

**Question 1**

Buenos días. ¿Cómo ... usted?

- A. estás
- B. es
- C. estamos
- D. está

*Figure 4.9 Multiple choice training task example*

### 4.4.3.5 Processing Input

Figure 4.10 and 4.11 show examples of the processing input training tasks. This task involves two steps. In the first step, participants are provided with explicit explanations about the areas in which they will likely experience difficulties processing the target grammar and why. Participants are then presented with learning strategies for overcoming these difficulties. Participants’ understanding of these explanations of both difficulties and strategies was
checked by way of gap fill tasks in summary texts (see figure 4.10). In the second step, participants attempt a series of simple tasks in which finding the right answer involves employing the strategies previously presented to overcome the input processing difficulties relevant to that particular grammatical concept.

Watch the video below about the challenges in learning grammatical gender in Spanish. Then complete the paragraph that follows.

Grammatical Gender doesn't communicate meaning. If you get grammatical gender wrong, a Spanish speaker will hear your mistake, but will understand what you want to say. This means you might ignore learning grammatical gender.

Grammatical gender is based around nouns. Nouns in Spanish are often in the middle of sentences or phrases and because our brains prefer to process words at the beginning or the end, this can mean that the noun is hidden.

Watch the video below on the strategies for dealing with the challenges of learning grammatical gender in Spanish. Then complete the paragraph that follows.

If you pay attention to the Spanish you hear and see and try to find the patterns in the language, you will quickly start to notice how grammatical gender works. You can also ask teachers and even check grammar books for help finding these patterns.

If you look for the words that have to agree with the noun, you can find clues that make it easier to determine the grammatical gender of the hidden noun.
This task was derived from VanPatten and Cadierno (1993) and VanPatten (2002). It was an attempt to introduce explicit instruction into the training, but rather than presenting the rules of the target grammar, it offered strategies for overcoming difficulties in processing and analyzing L2 input. The task was thus designed to promote deductive reasoning in which explicitly taught strategies are employed in practice tasks (VanPatten, 2002). Language analysis as deductive reasoning was explained to participants as part of the forming of concepts and hypothesis stages of the model (see figure 4.5).

4.4.3.6 Google Translate Task

Figure 4.12 shows an example of the Google Translate training task, which is an equivalence matching task. Participants are presented with an L2 sentence and its translation as generated by Google Translate (https://translate.google.com). Participants need to indicate if the
translation is equivalent in terms of grammar, meaning, both grammar and meaning, or neither grammar nor meaning.

This task was completely novel and was inspired by my observation as a TESOL instructor that a considerable number of L2 learners in the initial stages often rely on overly simplistic translation strategies, e.g. word for word. This task was intended to encourage more in-depth analysis of the products of translation as an L2 learning strategy. The task was designed to promote reasoning. In other words, the task required that the translation be analysed and that this analysis involved both semantic and morphosyntactic considerations, and that these considerations be understood as interrelated factors. Language analysis as reasoning with translational processes was explained to participants as part of the hypothesis testing stage of the model (see figure 4.5).

**QUESTION 1**

¿A quién le gusta jugar al básquetbol?

Who likes to play basketball?

- Only meaning is translated acceptably.
- Only grammar is translated acceptably.
- Both meaning & grammar are translated acceptably.
- Neither meaning nor grammar is translated acceptably.

*Figure 4.12 Google Translate training task example*
4.4.3.7 Feature Descriptions

Figure 4.13 shows an example of one type of feature listing training task. Participants are presented with a variety of questions designed to make them think explicitly about the features of the grammatical concept. For example, questions could ask what are the most frequent forms of the target grammar, what relationships exist between concept members and non-members, and what are the similarities or differences between analogous concepts in English grammar.

This task was inspired by Brown’s (2005) review of early research into concepts and categorisation. In particular, the current study’s feature listing training tasks were designed to encourage participants to formulate their own rules from a grammatical-knowledge-as-concept perspective. In addition, the task was also designed to promote reasoning of category membership and non-membership rules, and deductive reasoning of explicit knowledge of category membership (Murphy, 2004). This task was explained to participants as part of the clarification (or restructuring) stage of the model (see figure 4.5), particularly as the reflective reasoning encouraged by the task could lead to explicit formulations of the L2 grammatical system that could then be applied retrospectively to unanalysed acquired L2 knowledge.
4.4.3.8 Q&A

Figure 4.14 shows an example of the Q&A training task. This is a sentence reconstruction task, in which participants are asked a question with a visual prompt. They then need to unscramble a sentence by selecting words from a drop down menu for a given number of 'word' slots. The final sentence needs to answer the question grammatically correctly.

This task was inspired by problem-solving approaches to L2 pedagogy (e.g. Fotos & Ellis, 1991) that raise awareness and engage learners beyond the simple application of a rule to a single instance. The Q&A training task requires participants to construct the correct answer by means of more in-depth thinking than simply selecting the appropriate response. The task was designed to promote problem solving by eliciting the identification of the nature of the problem and the subsequent structuring a solution (Reed, 2000). It was also designed to promote analogical reasoning from the grammatical pattern in the question to the sentence
(Carroll, 1990). This task was explained to participants as part of both the hypothesis testing and the clarification (or restructuring) stage of the model (see figure 4.5).

**Figure 4.14 Q&A training task example**

4.4.3.9 Summary of training tasks

Table 4.5 shows each training task by type with a brief description of each task. Each task type is also matched with a specific stage of the language analysis model used by the current study (see section 3.2.4), as well as with the specific language analytic abilities the task type was designed to train. Information is also given about how many times each task was presented in the training course. The table is structured to reflect the relative importance of these tasks, which is indicated by the number of times a task type was presented in the course (see the final column). It should be noted that both the pattern matching and GJT tasks were intended to be presented in the training more times than they actually were. For an unknown reason, most likely a technical glitch in the LMS, one instance for each task type was not
made available to participants. The asterisked number in parentheses indicates the number of actual times these task types were presented to participants.

While all these task types reflect language analytic abilities that may be important in L2 learning, the relative importance placed on different task types was intended to relate to participants’ stages of L2 learning. As all participants were absolute beginners in L2 Spanish, noticing and pattern identification (Skehan 2002, see section 2.1.3) were regarded as the language analytic abilities most relevant to their L2 learning. In addition, L1 transfer, involving analogical reasoning, was also assumed to be an important language analytic ability at the initial stage of L2 learning. Consequently, task types that targeted noticing and analogical reasoning were privileged throughout the training. However, despite this emphasis, the training included a range of training task types that targeted each stage of language analysis and targeted a range of specific language analytic abilities.
### Table 4.4 Summary table of training task types and their relationship to language analytic abilities\(^{19}\)

<table>
<thead>
<tr>
<th>Task Type</th>
<th>Description</th>
<th>Targeted Area of LAAs</th>
<th>Targeted Ability</th>
<th>Instances in training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern Matching</td>
<td>First a grammatical pattern, presented in both English and Spanish, has to be identified and learned from exemplars highlighted in sentences. Then the same grammatical pattern has to be located in novel sentences in Spanish and written. English sentences are also presented in the question sentences to minimise lexical problems.</td>
<td>Construction stage of Awareness</td>
<td>Noticing, Analogy</td>
<td>5(*4)</td>
</tr>
<tr>
<td>Problem Solving</td>
<td>Given a small sample set of language in Spanish with translations into English, novel language must be made by recombining the grammatical patterns from the input.</td>
<td>Hypothesis Testing</td>
<td>Probability Solving, Inductive reasoning, Analogical reasoning</td>
<td>4</td>
</tr>
<tr>
<td>GJT</td>
<td>A sentence in Spanish is presented and must be judged to be grammatically acceptable (Yes) or unacceptable (No). Task has a time limit.</td>
<td>Forming Concepts</td>
<td>Recognition, Noticing</td>
<td>4(*3)</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>An incomplete sentence in Spanish must be completed by selecting the most appropriate of four possible answers.</td>
<td>Reconstruction stage of Clarification</td>
<td>Problem Solving, Reasoning (all types)</td>
<td>3</td>
</tr>
<tr>
<td>Processing Input</td>
<td>Processing input involves two steps. In the first step, explicit explanations are provided to participants about the areas in which they are likely to experience difficulties processing grammar and why. Strategies for overcoming these difficulties are also presented. In the second step, participants undertake a series of simple tasks in which finding the right answer involves overcoming the input processing difficulties relevant to that</td>
<td>Forming Concepts and Hypothesis Formation</td>
<td>Deductive reasoning</td>
<td>2-3</td>
</tr>
</tbody>
</table>

\(^{19}\) See table 4.6 for schedule of tasks in language analytic abilities training course.
This is an equivalence matching task. An L2 sentence is presented with a translation generated by Google Translate. Trainees need to indicate if the translation is equivalent in terms of grammar, meaning, grammar and meaning, or neither grammar nor meaning.

Feature Description
Participants are presented with a variety of questions designed to make them think explicitly about the features of the grammatical concept, for example, most frequent forms, concept members and non-members, and differences between analogous concepts in English grammar.

Q&A
This is a sentence reconstruction task, in which participants are asked a question with a visual prompt. They then need to unscramble a sentence to answer the question grammatically accurately.

4.4.4 Explainers
As mentioned, the training in LAAs involved not only practice tasks but also instruction in how language can be analysed for the purposes of L2 learning. As part of the training, every week began with short instructional videos, or explainers, about particular aspects of language analysis in L2 learning, e.g. L1-L2 analogies. The following section describes the design and content of the explainers in the LAAs training.

4.4.4.1 Design of explainers in the training
The training was designed to achieve three objectives. First, it aimed to raise participants’ awareness of the many ways in which L2 input can be analysed for the purpose of learning the patterns and structures of the L2. This objective was based on Snow and Lohman’s (1984)
review of studies that had attempted to train (general) aptitude. In some of these studies, participants’ post-training test scores actually decreased. Snow and Lohman speculated that in these cases, participants who had pre-existing and functional approaches to learning that drew on their individual strengths had been trained in new approaches that were less effective or disruptive and thus scored worse on the post-test. To avoid this effect, the current study adopted a more open approach to training with less pressure to abandon pre-existing, functional learning approaches, particularly those that drew on aptitudinal strengths. Participants were not forced to adopt one single new skill or strategy, but were instead offered a range of skills and strategies and encouraged to experiment with them all, and to adopt those that they found effective or useful. Put otherwise, the training aimed to increase the range of strategies with which participants analysed L2 input for learning, not to train them in a single ‘superior’ strategy.

The second objective of the training was to offer the participants multiple opportunities to experiment, practice, and apply these various language analysis strategies. However, the training did not compel participants to complete all of the training, partly due to ethical considerations, but also to avoid replicating the problems of low motivation and animosity towards the training experienced in the Politzer and Weiss (1969) aptitude training study. Instead, the training was designed to be as relevant as possible to the participants’ coursework.

The third objective of the training was to encourage the application of language analytic abilities to real L2 learning. To make the training as relevant as possible to the participants’ L2 Spanish course, the vast majority of the words and phrases presented in the training course were taken from the participants’ course textbook, as was the majority of the target
grammatical concepts. Where possible, the presentation of grammatical concepts in the training was scheduled to complement required grammar knowledge in the L2 Spanish course. This incorporation of the course material into the training was designed to encourage and stimulate participants to apply their new language analytic abilities to their learning in their L2 Spanish course.

These three objectives were achieved in a variety of ways. The second and third objectives were achieved mainly through the structuring of the training course, e.g. by the considered scheduling of grammatical concepts. However, the first objective was mainly achieved through graphic design features, e.g. enhanced text (Jourdenais, Ota, Stauffer, Boyson, & Doughty, 1995), and content presentation (see figure 4.6, repeated here as 4.15). The most important aspect of the content for raising awareness of abilities for language analysis was the video explainers presented at the beginning of each week of the training.

**Figure 4.15 Example of text enhancement in training task (from figure 4.6)**
The training for LAA was based on a proposed model developed as part of the current study (see section 3.2.4). As described in Chapter 3, this model was based on general learning and involved five stages. These five stages were used to structure the training program, with one stage being the focus of each week. Each week of the training commenced with a short video that introduced the current stage of the model and the abilities most implicated at that stage.

4.4.4.2 Video explainers

The design for each of the video explainers followed the same basic structure. Each video consisted of a simple animation covering one of the five stages of language analysis from the proposed model. Each stage was explained by way of a simple fictional character, Aal (LAA spelled backwards). Aal exemplified L2 learning by way of language analysis; in each video, Aal gave concrete examples of how he used his language analytic abilities to learn another language, in this instance, Spanish.

Each video was short, approximately two to three minutes in length. They were narrated and created by the researcher using the free online animation program, PowToon (https://www.powtoon.com/). The animations were simple and dynamic, and incorporated textual enhancements to highlight, reinforce, and extend the narrated explanations. Each video also finished with a checklist of all abilities explained or mentioned to consolidate the main points of the video. Participants were then given a simple gap-fill summary task to check they had understood the content and to revise the key points. In total, five video

---

20 To see the video on awareness in LAAs and L2 learning as an example of the video explainers, follow the link: https://www.youtube.com/watch?v=B7lerK9gL8&list=PLVjiujiik9D7tqnMFJ4hHlZ2Qq0DzWRZug&index=2
explainers were created, listed below with the stage of the LAA model that they explained and the title of the video presented to participants.

1. Awareness: What is awareness?
2. Concept formation: How do you form concepts?
3. Hypothesis formation: What do you do with concepts?
4. Hypothesis testing: Making and learning from mistakes
5. Concept clarification/restructuring: How to keep an open mind

In short, the video explainers were intended to engage and prepare the participants for the training tasks, and to build among participants an expectation that these training tasks would be novel and challenging. The video explainers were also intended to demonstrate the underlying purpose of LAAs and to offer a broader overview of the language learning process, an understanding that might be important in developing L2 aptitude abilities in general (McLaughlin 1995, Sternberg 2002, see section 3.1.3.4. We now detail the content of the training tasks themselves, namely the Spanish linguistic structures presented to participants for analysis.

### 4.4.5 Linguistic structures

As part of the training course in the study, six grammatical concepts of Spanish were included:

- *estar* (*to be*) (section 4.4.5.1)
- grammatical gender (section 4.4.5.2)
- number agreement (section 4.4.5.3)
- noun-adjective placement (section 4.4.5.4)
These six concepts were chosen for four main reasons. First, they were identified as being problematic for learners of L2 Spanish. Second, due to their difficulty, they pose challenging and real problems for linguistic analysis. In other words, because these aspects of Spanish are unlikely to be understood quickly and simply, they encourage more in-depth thinking, i.e. language analysis. Third, the nature of the grammatical concepts offers opportunity to develop specific aspects of language analysis, e.g. recognition (of word forms), noticing (of patterns in word order and morphology), analogical reasoning (of L1-L2 similarities and contrasts), inductive reasoning (of patterns in word order and morphology), deductive reasoning (using exemplars and prototypical structures), and problem solving. Importantly, though, participants were not expected to master these grammatical concepts at this stage but rather to learn to analyse the L2 input in which instances of these concepts could be found. Evidence of learning to analyse L2 input would be increased accuracy in the training tasks; no claim to measure L2 achievement is thereby being made. Finally, all of these grammatical concepts except one (the animate ‘a’) were present in the classroom course for Spanish, so were relevant to participants’ L2 learning.

To select these six particular grammatical concepts, discussions were held amongst the research team (which included two highly-experienced L2 Spanish teachers, one of whom is a L1 Spanish speaker), the grammar topics from the participants’ course textbook were examined, and literature on L2 Spanish grammar (e.g. Whitley, 1986) was consulted.

Before detailing what was presented for each of the six grammatical concepts of Spanish, a brief explanation of the overall aspects of the linguistic structures included in the training

- *gustar* (to like) (section 4.4.5.5)
- the animate ‘a’ (section 4.4.5.6)
course is required. Rather than a traditional pedagogical approach involving the deductive instruction of grammatical rules, the training conceived of grammatical knowledge from a more general cognitive perspective. In particular, grammatical knowledge was viewed in the same way that any knowledge is, namely as concepts (Murphy, 2004). On such a construal, the key consideration in designing the training content was: What do people learn through analysing language? Adopting Murphy's (2004) view of mental representations as concepts, knowledge of grammar was construed as consisting of exemplar words, phrases, and sentences that the learner remembers (although not perfectly), organized by relationships (e.g. similarity, function, context, etc.), with prototypes emerging as those parts that best typify the concept in the mind of the learner (Murphy, 2004). In this context, ‘exemplars’ refers to those actual tokens or instances that typify a given concept, and ‘prototypes’ refers to an abstraction of the tokens that comes to represent a type. For example, an exemplar of the Spanish verb estar might be ‘estoy estudiando’ (I am studying), while the prototype might be estar + verb-ndo (to be + verb-ing).

In general, the training course focussed on four main aspects of grammatical knowledge across all six concepts of Spanish grammar chosen, namely, word order, agreement, word forms, and parts of speech. In addition, the training also engaged related areas of processing, namely, visual processing or recognition of words, semantic vs. grammatical processing, and challenges arising from English-Spanish similarities and contrasts. These details varied in importance and prominence across the six target grammatical concepts. In the following section, we briefly cover the specific details of each target concept in the training.

For each grammatical concept, all instances were identified in the textbook, Dos Mundos 7th ed. (Terrell et al., 2010), and collated into a small corpus. Where the textbook contained
insufficient examples of a grammatical concept, Spanish language training materials from the Foreign Service Institute (https://fsi-languages.vojik.eu/) were used to supplement the corpus.

4.4.5.1 Estar (To be)

Spanish has two copular verbs, *ser* and *estar*, which roughly translate as *to be*. While the two verbs exist in complementary distribution, when it comes to adjectives their distributions overlap, which, in addition to the perfective/imperfective contrast, is the main source of confusion for L1 English learners of L2 Spanish (VanPatten, 2010). This confusion requires more analysis on behalf of the L2 Spanish learner and is the reason this grammatical concept was chosen for the training. *Estar* is normally acquired after *ser* and with more difficulty (VanPatten, 2010), so was selected as the target grammar.

Exemplars of *estar* were divided into two types: verbal constructions and verbal forms (i.e. conjugations). Following VanPatten (2010), three verbal constructions of *estar* were identified to be included in the training. These three constructions, in order of prototypicality, are:

1. *estar* + verb-*ndo* (progressive aspect) (e.g. *El peluquero está cortando el pelo*. The hairdresser is cutting the hair.)
2. *estar* + true locative (e.g. *Está en la clase*. She/he’s in the classroom.)
3. *estar* + adjective (to express conditions) (e.g. *Luis está enfermo*. Luis is sick.)

Prototypicality for the forms of *estar* were deduced from their frequency in the training corpus. The forms in order of prototypicality are:

1. *está* (third person singular, second person singular polite)
2. *estoy* (first person singular)
3. estás (second person singular informal)
4. están (third person plural)
5. estamos (first person plural)

Borderline cases for estar were also divided into two groups. Visual distractors were those words that had similar visual forms to those of estar, i.e., eso, esto, esta, este, ese, esa, esos, estos, esas, éste, ésta, and ésa. Confusions between ser and estar were also included.

VanPatten (2010) argues that ser is acquired first for all instances of ser and estar, with the learner acquiring estar + verb-ndo first and then estar + true locative later. For this reason, the training included marked instances of the estar constructions with ser.

The objective of the training tasks with estar was to make participants analyse the L2 input in such a way as to become aware of the key aspects of the grammatical concept. These key aspects included the existence of the three verbal constructions and the prototypicality of the está conjugation (as detailed above), as well as word order patterns and diacritic use. Certain tasks also encouraged conscious reflection on the grammatical concept and explicitly asked about English-Spanish analogies and visual distractors that emphasised the differences in grammatical function, i.e. estar is a verb and the distractors were determiners. Another difference presented between estar and its visual distractors was to highlight the placement of diacritics, i.e. the accent marker that falls on the final syllable of certain present indicative conjugations of estar, e.g., está and estás, as opposed to instances which include no written accent (e.g., esta silla) or where the accent falls in the first syllable, as in determiners without nominals (ésta [silla]). While these aspects of estar are not dissimilar to what was presented in the grammar exercises of the class textbook, the key difference resided in how the knowledge was presented to the learner. LAAs emphasise bottom-up learning and the learner...
‘figuring out’ the L2 system for themselves as opposed to top-down rule-learning. The training tasks for *estar* encouraged participants to construct their own understanding of the grammatical concept with an emphasis on acquiring exemplars and identifying grammatical patterns (as opposed to rules).

4.4.5.2 Grammatical gender

Spanish marks nouns for grammatical gender, either masculine or feminine. Agreement rules between nouns and modifiers are also present, with adjectives and determiners modifying nouns also marked for grammatical gender. Gender marking is ubiquitous and arbitrary, which makes it problematic and unavoidable for L2 learners. In addition, grammatical gender is almost absent in English, the L1 for the majority of participants, making the concept of grammatical gender novel and lacking direct L1-L2 transfer or obvious L1-L2 analogies. These challenges offer good opportunities for analysis of L2 input, and were the main reason grammatical gender was included in the training.

Exemplars of grammatical gender were taken from Whitley (1986), who details the relative frequencies of the most common grammatical gender morphology (see table 4.6) but who also includes borderline, or less frequent, cases (not included in table 4.6). Knowledge of grammatical gender (in contrast to biological gender) as a concept was explicitly presented to participants. Part of speech information was also included in grammatical gender, as determiners and adjectives of the same noun phrase must all agree in gender with the head noun in Spanish.
Table 4.5 Grammatical gender of nouns in Spanish (Whitley, 1986)

<table>
<thead>
<tr>
<th>Masculine nouns</th>
<th>Suffixes and frequency rate</th>
<th>Feminine nouns(^{21})</th>
<th>Suffixes and frequency rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>pelo</td>
<td>-o  (99.7%)</td>
<td>blusa</td>
<td>-a  (98.9%)</td>
</tr>
<tr>
<td>suéter</td>
<td>-r  (99.2%)</td>
<td>civilización</td>
<td>-ción/-sión (100%)</td>
</tr>
<tr>
<td>salón</td>
<td>-n  (96.3%)</td>
<td>[sinopsis]</td>
<td>-sis/-tis (99.2%)</td>
</tr>
<tr>
<td>país</td>
<td>-s  (92.7%)</td>
<td>ciudad</td>
<td>-d  (97%)</td>
</tr>
<tr>
<td>nombre</td>
<td>-e  (89.2%)</td>
<td>[cárcel]</td>
<td>-l  (96.6%)</td>
</tr>
</tbody>
</table>

Although L2 aptitude tests of LAAs focus on bottom-up inductive language learning skills, deductive reasoning may also be common in L2 learning and may even enhance performance on tests of LAAs, e.g. MLAT IV Words in Sentences (Carroll, 1962). Consequently, much of the grammatical knowledge for grammatical gender was made explicit in the training for two reasons. First, as previously mentioned, this grammatical concept is almost completely absent in English. Thus, the difficulty of inferring patterns of grammatical gender may be overly challenging for inexperienced learners to infer from limited L2 input and potentially even demotivating if the concept remained opaque to the student. So explicit knowledge of this concept was seen as enabling learning and keeping students more motivated. Second, to take advantage of this emphasis on explicit knowledge of grammatical gender, processing instruction style tasks were used in the training. Processing instruction includes explicit explanation of troublesome areas for learners and then suggests strategies to address these troublesome areas (VanPatten, 2002). This approach was therefore deemed most suitable for the grammatical concept of grammatical gender.

\(^{21}\) Spanish nouns inside square brackets were not present in the participants course textbook, so were displayed for participants but were not included in any of the training content.
4.4.5.3 Number agreement

Number agreement in Spanish was chosen for two reasons. First, number agreement in Spanish is more complex because it affects adjectives, unlike English. Second, this extension of a grammatical concept that most learners would at least be familiar with offered a rich opportunity for analogical and inductive reasoning. In addition, the fairly regular repetition of morphological marking of number agreement across word classes also offered a rich opportunity for recognition, noticing, and inference of patterns.

Exemplars for number agreement were taken from the course textbook, Dos Mundos 7th ed. (Terrell et al., 2010). Plural forms were presented in order of typicality as follows:

1. vowel + -s
2. consonant + -es
3. z + /c-es

Exemplars of agreement were presented in order of typicality as follows:

1. determiner + noun (las mujeres, the women – las is the feminine plural definite article in Spanish)
2. noun + adjective (orejas grandes, big ears – translated literally as ‘ears big’ with big taking the plural form)
3. subject + verb (Ellos estudian en español, they study in Spanish – estudian is the third person plural form of the verb)

4.4.5.4 Noun-adjective placement

The reasons for including this grammatical concept in the training were essentially the same as for number agreement. English has possible analogs of this Spanish grammar pattern, but
the noun-adjective placements in the two languages differ. The fairly consistent syntactic pattern also offered good opportunities for recognition, noticing, and inference.

Exemplars of noun-adjective placement were taken from the course textbook (Terrell et al., 2010). Noun-adjectives were presented in order of typicality:

1. Noun + single adjective (*vestido nuevo*, new dress – translated literally as ‘dress new’)
2. Noun + multiple adjectives (*vestido rojo muy bonito*, very beautiful red dress – translated literally as ‘dress red very beautiful’)

Determiners were also introduced in the examples to provide a more complete picture of the grammatical pattern.

4.4.5.5 *Gustar (To like)*

This grammatical concept was chosen for its novelty, particularly to L1 English learners. Specifically, the *gustar* construction offered an excellent opportunity to extend analogical reasoning due to the grammar-semantic mismatch – *gustar* is semantically equivalent to the English verb ‘someone likes something’, but grammatically *gustar* functions more like the English verb ‘something pleases someone’. Therefore, the grammatical concept of *gustar* was used as an opportunity to explore difficulties with translation when linguistically analysing the L2 input.

Exemplars of *gustar* were taken from the course textbook (Terrell et al., 2010). Exemplar constructions were presented in order of prototypicality as follows:

1. *gustar* + postverbal nominative subject (*me gusta la nieve*, I like the snow – translated more literally as ‘the snow pleases me’)

---

22 The explanations of *gustar* follow the pedagogical grammar of the textbook, not a descriptive grammar.
2. *gustar* + verbal subject (*me gusta esquiar*, I like to ski – translated more literally as ‘to ski pleases me’)

Exemplar forms were presented in order of prototypicality as follows:

1. *me gusta*
2. *te gusta*
3. *nos gusta*
4. (*me/te/nos*) *gustan*
5. (*me/te/nos*) *gustas*

The borderline case of *me gusta ir a la playa* (I like to go to the beach) was also presented for the English ambiguity of ‘to’ when glossing/ translating to Spanish. This ambiguity makes categorising the grammatical patterns of this construction a challenging task of learning that ‘to’ in English can take multiple syntactic categories. In particular, this example shows the dual use of ‘to’ in English: (1) as a particle in the infinitive form of the verb (e.g. to go) and (2) as a preposition indicating movement (e.g. to the beach). In contrast, Spanish does not show this same grammatical pattern. The *-ar/-er/-ir* infinitival ending of the verb in Spanish replaces the infinitival particle *to*. Although the Spanish preposition indicating movement (*a*, ‘to’) has other functions in other constructions (see section 4.4.5.6 for more details), only the preposition of movement function of *a* (*to*) was used in the training tasks for the verb *gustar*.

### 4.4.5.6 The animate 'a’

This grammatical concept of the animate 'a’ was included because it is redundant from a semantic processing perspective yet obligatory from a syntactic perspective. Thus, the
animate ‘a’ is a subtle aspect of Spanish grammar that poses a long-term challenge to learners.

Exemplars for the animate ‘a’ were developed in consultation with the research team and taken from the course textbook, *Dos Mundos 7th ed.* (Terrell et al., 2010). Exemplar constructions were presented in order of prototypicality as follows:

1. subject + verb + ‘a’ + direct object (*Ernesto no conoce a muchas personas*, Ernesto does not know many people, cf. *Ernesto no conoce muchas ciudades* – without animacy marker)

2. subject + verb + ‘a’ + direct object + indirect object (*le escribe una carta a su amiga*, she is writing a letter to her friend)

Borderline cases that highlighted the difference between the animate 'a’ and 'a’ as a preposition of movement were also included. For example, in the sentence *La gente va a la plaza para sentarse, conversar o simplemente para mirar a las personas que pasan* (The people go to the plaza to sit, chat, or simply to watch the people passing by), the first instance of ‘a’ in *a la playa* is equivalent to the English preposition of movement ‘to’ in ‘to the beach’. However, the second instance of ‘a’ in *mirar a las personas* has no equivalent in English ‘watch the people’. This contrast offered more opportunities for participants to not only analyse new patterns in the L2 input but also to reconsider previous hypotheses that may have been incomplete.

### 4.4.6 Syllabus for the language analytic abilities training course

Table 4.7 details the syllabus of the training course used in the current study. As mentioned, each week of the training course included three components. The primary component was the
explanation of one aspect of language analysis in L2 learning (LAAs explainers). As
described above (section 4.3) each week participants were first presented with a video that
explained the five stages of the LAA model developed for the purposes of training (the week
6 explainer was a summary overview of all five stages). The secondary component was the
training tasks, normally five a week except for weeks three and four. In week three,
participants were introduced to a new task type, problem solving. Since these tasks were
more complex and demanding than the other tasks, more time was given to these tasks at the
beginning while participants familiarized themselves with the problem-solving task type. The
training tasks were an opportunity to apply the strategies and ideas from the explainers to real
problems of L2 analysis. Subsequently, training tasks for each week were selected for their
relevance to the stage of LAA that was the focus for the week, while still offering practice in
abilities that were relevant to other stages in the model. For example, the focus of week one
was the awareness stage of the model, which was best represented at this stage by the pattern
matching task. Week one also included the feature description task, which was considered
more relevant to the clarification/restructuring stage of the LAA model. However, the explicit
questioning of the L2 grammatical concept in the feature description task was also considered
to raise learner awareness and to encourage noticing of more features of the concept. For this
reason, it was also included in the week one tasks. The third component of the training was
the L2 input used for analysis in the training tasks. The L2 Spanish grammatical concepts
also served two other purposes. First, the integration of training content with class and
textbook content was intended to make the training relevant to participants’ L2 learning and
thus increase motivation to engage with the training course. Second, the acquisition of factual
knowledge was identified in the literature as important in successful problem solving (Reed,
2000) and considered an important part of the L2 learning experience. On this basis, rather
than use disparate data sets of various grammatical concepts in various L2 languages, a cohesive course of L2 grammatical concepts from one language\textsuperscript{23} was considered to be a better choice for the acquisition of L2 knowledge.

In summary, each week of the training course involved the explanation of strategies and metacognitive knowledge with opportunities to apply and practice such knowledge with L2 input that was both relevant and useful to participants.

\textsuperscript{23} The exception to this rule was the first instance of the problem-solving tasks, which involved using past problems from the OzCLO (past problem sets can be found at http://ozclo.org.au/past-problems/)
4.5 Data collection

Data collection was carried out completely online over a 9 week period. The L2 aptitude pre-testing took place over weeks 1 and 2, the training course for language analytical abilities over weeks 2 to 7, and the L2 aptitude post-testing over weeks 8 and 9.
Data collected included training course completion, time on task, relative achievement on the training course, and scores across all L2 aptitude scores for both pre- and post-tests. Course completion was operationalised as the number of training tasks attempted at least once. Relative achievement on the training scores was calculated for both scores and time taken on the first attempt of each training task attempted. An average of the group was calculated for both measures and then each participant’s performance was calculated against the average (see section 5.2.1.4 for more details). L2 aptitude was operationalized as the test scores for each sub-test of the LLAMA test (see 4.4.1.2 for more details) as calculated by the LLAMA test program (see Meara, 2005).

4.6 Analysis

To answer research question 1.a. (see section 3.3), a paired-samples t-test was employed to discover any differences between pre- and post-test scores for all LLAMA tests. A two-sided t-test was employed, as previous research into aptitude training has shown that training can actually inhibit individual L2 aptitude abilities (Snow & Lohman, 1984, see section 3.1.3.4 for discussion). For this reason, although the study is primarily concerned with increases in L2 aptitude test scores – which implies a one-sided t-test – a two-sided t-test was used to uncover any inhibitory effects of training or L2 learning experience. Paired-samples t-tests were also carried out for two participant groupings – low (below average) and high (above average) pre-test scores. For each test individually, participants who scored below the average level as per the LLAMA test manual (Meara, 2005) were placed in the low group and

---

24 An attempt is operationalized as a task registered by the LMS as having been completed. For this to occur, the participant needed to click on the “submit” button, presumably after having completed the entire task.

25 ‘Below’ and ‘above’ average test scores was determined according to interpretations given in the LLAMA test manual (Meara, 2005). In other words, test scores that were considered ‘average’ or ‘very poor’ were categorised as low and test scores that were considered ‘good’ and ‘outstandingly good’ were categorised as high.
those who scored above average were placed in the high group. This analysis was carried out to check if any differences measured were consistent across all participants. Pearson’s r was also calculated between pre- and post-test scores for each LLAMA test. This was done to check the test-retest reliability.

To answer research question 1.b. (see section 3.3), a random-forest regression model was employed to identify the relative importance of explanatory variables and their contribution to predicting the response variable, in this instance, the LLAMA F post-test scores.

Random forest regression models are a non-parametric alternative to general linear models, especially when there are many potential interactions among variables and the sample size is small but the number of predictors is large (Levshina, 2015). Random forest regression models are grown from conditional inference, or classification, trees. Classification trees are made from binary recursive partitioning. Put otherwise, predictor variables are tested against the response variable looking for the predictor with the strongest relationship. The data then undergoes a binary split by that variable, e.g. in the case of a test score, the data may be split into scores from 0-40% and 41-100%. This process is then repeated until all options have been exhausted. (For more information or explanation, see Liaw & Wiener, 2002; Levshina 2015; Strobl et al., 2009.)

Liaw and Wiener (2002) describe the process of constructing a random forest model as follows. A specified number of bootstrapped samples are taken from the dataset. For each sample, a classification tree is made. At each branch of the classification tree a specified number of predictors is randomly sampled and the best split among those predictors is chosen. The model prediction is made by aggregating the predictions among the classification trees. An error estimate is also calculated in two steps. For each tree, the data included in the
sample is used to predict the data not in the sample. Each error rate is then aggregated across all trees, creating a total error estimate for the random forest model (see section 5.2.2 for details).

4.7 Limitations

Because the training course engaged actual learning behaviours outside of the laboratory, self-selection was found to be an important limiting factor the design of this study. As discussed above, participants were not compelled to complete pre-testing of L2 aptitude, language analytic abilities training, or post-testing of L2 aptitude. As the study was conducted over the first nine weeks of the university semester, participants’ completion of the training tasks in particular was less than consistent. While this challenges the internal validity of the study, it also increases its ecological validity. In other words, the current study is designed to more accurately reflect the real learning of the sample population of university students.
CHAPTER 5: ANALYSIS AND RESULTS

This chapter details the results of the statistical analyses carried out to answer the research questions and is organised as follows.

Section 5.1 details the analyses and results for research question 1.a., that is, Are pre- and post-test scores on the LLAMA tests significantly different? This question was investigated using three different approaches: paired-sample $t$-tests at the cohort level, paired-sample $t$-tests at the sub-group level (i.e. above-average vs below-average pre-test scores), and pre- and post-test correlations. For each approach, all analyses are first explained and their use justified. Before reporting the results for the inferential statistics, the flow of participant numbers is detailed and the data are checked to ensure all assumptions of the relevant tests are met. Possible differences between participants based on background characteristics are also checked and descriptive statistics for all LLAMA tests are provided.

Section 5.2 details the analyses and results for research question 1.b, that is, Do training factors explain more variance in LLAMA F post-test scores than L2 aptitude variables? This question was investigated using two different approaches: a random forest regression model and a stepwise linear regression. For each approach, all analyses are explained and justified. Before reporting the results for the inferential statistics, the data preparation for all independent variables is explained in detail.

All statistical analyses were carried out in R 3.3.2 (R Core Team, 2016), using an alpha level of 0.05 for all relevant tests (see Appendix C for additional details of analyses).
5.1 Research question 1.a: Are post-test scores for L2 aptitude, operationalised as the LLAMA tests, significantly different to pre-test scores?

The first research question focussed on the stability of L2 aptitude measures, operationalised as the LLAMA tests. A significant increase in the LLAMA tests would suggest that L2 aptitude is not as stable as traditionally assumed, but in fact can be enhanced through experience and possibly training. Of the four tests, the measure of language analytic abilities, operationalised as the LLAMA F, is of particular importance, as the second research question is only relevant if significant gains in the LLAMA F are present. Parametric paired-samples t-tests were carried out to find any significant differences between participants’ pre- and post-test scores for the LLAMA B, D, E, F, and Total26 tests. To quantify the size of any possible gains, Cohen’s $d$ was also calculated.

Before detailing the results for the paired-samples t-tests, details are provided regarding the flow of participant numbers during data collection (section 5.1.1), assumptions of paired-samples t-tests (section 5.1.2), and checks conducted for any confounding effects of background characteristics on participant pre-test results (section 5.1.3). Descriptive statistics for all LLAMA tests are also provided (section 5.1.4). Section 5.1.5 reports the results of the paired-samples t-tests at the group level followed by section 5.1.6, which reports the results for the paired-samples t-tests at the sub-group level (below- vs above-average pre-test scores). Section 5.1.7 reports the results of the correlational analysis of pre- and post-test LLAMA test scores. These findings are then summarised in section 5.1.8.

26 “Total” was an aggregate score of LLAMA B, D, E, and F tests for an individual participant and not a separate test.
5.1.1 Participant flow and recruitment

Participants were recruited between 02 and 15 March 2015. All testing was administered over the Internet, with pre-testing completed between 09 and 22 March 2015 and post-testing between 27 April and 03 May 2015. Of the 128 students who initially agreed to take part in the study, 115 completed all four of the LLAMA pre-tests. Of the 115 students who completed all the LLAMA pre-tests, 90 completed post-tests for the LLAMA B, D, and F, and 85 completed the LLAMA E post-test. Paired-samples $t$-tests were therefore carried out for each LLAMA test separately.

5.1.2 Assumptions of paired-samples $t$-tests

Parametric paired-samples $t$-tests were carried out to check if the difference between pre- and post-test scores from the same participants was statistically significant. $T$-tests were two-tailed to factor in the likelihood that some participants would experience negative gains as a result of an inhibitory effect of instruction, either from the L2 Spanish course or the language analytic abilities training course (see section 3.1.3.4 for discussion). The assumptions of paired-samples $t$-tests are that gain scores between tests are normally distributed and there are no outliers. Table 5.1 shows the numerical tests of normality for each LLAMA test: skewness, kurtosis, and the results for Shapiro-Wilks goodness-of-fit test.
Table 5.1 Skewness, kurtosis, and Shapiro-Wilks for LLAMA tests

<table>
<thead>
<tr>
<th>LLAMA tests</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>W (Shapiro-Wilks)</th>
<th>p (Shapiro-Wilks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.41</td>
<td>0.86</td>
<td>0.98</td>
<td>0.18</td>
</tr>
<tr>
<td>D</td>
<td>0.23</td>
<td>0.73</td>
<td>0.97</td>
<td>0.02</td>
</tr>
<tr>
<td>E</td>
<td>-1.06</td>
<td>3.22</td>
<td>0.88</td>
<td>0.00</td>
</tr>
<tr>
<td>F</td>
<td>-0.41</td>
<td>-0.18</td>
<td>0.97</td>
<td>0.07</td>
</tr>
<tr>
<td>Total</td>
<td>1.06</td>
<td>2.55</td>
<td>0.93</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Skewness is a measure of the symmetry of scores distributed around the mean (the length of the tail in a histogram), with levels between +/- 2 indicating a normal distribution (Weinberg & Abramowitz 2002:278). All the LLAMA tests fall well within this range of acceptability.

Kurtosis is a measure of the concentration of scores around the mean (the “peakiness” of the histogram), with levels between +/- 3 indicating a normal distribution (Crawley 2013:353). Only the LLAMA E falls (just) outside of this range of acceptability. Shapiro-Wilks goodness-of-fit test is a formal test of normality with non-significant \( p \) levels indicating a normal distribution (Levshina 2015:56). Only the LLAMA B and LLAMA F have non-significant \( p \) values, indicating that the LLAMA D, E, and Total scores are not normally distributed. Standard deviations showed that variance between pre- and post-tests for LLAMA E were also not quite homogeneous (see table 5.3).

Visual checks for normality for all LLAMA tests include histograms in figures 5.1 and 5.2, QQ (quantile-quantile) plots in figures 5.3 and 5.4, and box plots in figures 5.5 and 5.6.
The histograms and QQ plots confirm the non-normality indicated in the numerical checks. Examining the boxplots for each test, there also appear to be some outliers in LLAMA B, D, E, and Total.

Figure 5.1 Histograms of LLAMA tests B, D, E, and F
Visual checks of the histograms indicate that the LLAMA E, LLAMA F, and LLAMA Total are not normally distributed, but the deviations from normality are not large. This can be cross-checked with the QQ-plots, which show that most of the LLAMA tests seem to have longer tails, but again, the deviations from normality do not appear substantial.
Figure 5.3 QQ plots for LLAMA tests B, D, E, F

Figure 5.4 QQ plot for LLAMA Total
Figure 5.5 Boxplots for LLAMA B, D, E, F

Figure 5.6 Boxplot for LLAMA Total
The boxplots indicate a small number of outliers in the LLAMA B, D, and Total. In addition, the LLAMA E shows a very restricted interquartile range (indicated by the top and bottom edges of the box) with a number of outliers on both sides of the distribution.

However, results from paired-samples $t$-tests are still considered valid with non-normal distributions, provided that the departures from non-normality are not too extreme. As the data from the current study were judged to contain only relatively small departures from normality, it was decided to continue with the planned parametric tests. Outliers were also retained in the data set, thereby maximising the sample size and retaining as much power in the statistical tests as possible. In addition, their impact on the results can be seen from scatter plots (see figures 5.9 and 5.10 in section 5.1.5), and hence can be factored into the interpretation of the analysis.

### 5.1.3 Background variables of participants

Before running any inferential statistical tests, an independent factorial ANOVA was run for each LLAMA test and LLAMA Total for age, L1, and L2 self-reported experience level (see section 4.2 for summary descriptions of each variable) to determine if significant differences existed between for these individual factors in the sample. L1 was a binary factor of English vs. non-English, which was relevant to the study, as English was the language of instruction and many participants were not L1 speakers of English. The L2 experience level factor attempted to capture experience in the process of learning second languages, as opposed to mere proficiency or number of languages learned, and consisted of four levels (see section 4.2 for descriptions of these levels).
No significant differences were found for LLAMA B, D, E, or LLAMA Total for age, L1, and L2 experience level. However, a significant effect of L2 experience level was observed for the LLAMA F, $F(3, 109) = 3.3, p = 0.023$. Stepwise Akaike’s information criterion (AIC; Crawley, 2013) was then carried out, which returned a minimal model with L1 and L2 experience level factors for the LLAMA F pretest. Tukey’s test of honest significant differences (Crawley, 2013) was then run to find which factors differed. Results showed that the difference in LLAMA F test scores for both English and non-English speaking participants was between those with extensive and those with some L2 experience levels (see table 5.2 for numbers). However, although this difference was evident, it was not possible to pursue these groupings in any of the subsequent analyses, as the number of participants who had extensive L2 experience – only eight in total – was too small.

Table 5.2 LLAMA F pre-test scores for English vs non-English L1 participants for differing levels of L2 experience

<table>
<thead>
<tr>
<th></th>
<th>None</th>
<th>Low</th>
<th>Some</th>
<th>Extensive</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>48.49</td>
<td>52.86</td>
<td>52.50</td>
<td>75.00</td>
</tr>
<tr>
<td>Other</td>
<td>NA</td>
<td>NA</td>
<td>38.57</td>
<td>62.00</td>
</tr>
</tbody>
</table>

In summary, independent factorial ANOVA showed that participants did not differ on any of the LLAMA pre-tests for age, L1, or L2 experience level, except for LLAMA F. L1 English and non-English speaking groups who had extensive L2 learning experience had significantly higher LLAMA F pre-test scores compared with those with less L2 learning experience. However, the number of participants with extensive L2 learning experience was too small (n=8) for these groups to be used in subsequent analyses. In other words, although this result appears to suggest that extensive experience leads to significant increases in LLAMA F test
scores, the number of participants across all categories of L2 experience did not permit further analyses either for research question 1.a. or 1.b. and so all participants were collapsed into one single group. The variable of L2 learning experience was not judged to impact analyses related to the training of language analytic abilities either, because the sample overall could be considered to have no or some L2 learning experience (see table 4.4 in chapter 4). As a final note, the fact that the L2 proficiency levels were self-reported and not explicitly elicited questions the reliability of this data for the cohort as a whole. This suggests that any analysis carried out with these self-reported data may not be a highly accurate reflection of the differences between subjects.

5.1.4 Descriptive statistics

Descriptive statistics were calculated for all tests. The mean, range, standard deviation, variance, and 95% confidence interval (CI) for each LLAMA pre- and post-test are reported in table 5.3.
In general, the LLAMA test score means and standard deviations align with Granena’s (2013) exploratory study, suggesting that the sample is acceptable. However, there are three points worth noting in the data. First, the range on each test is large. In most tests, the minimum score is 0/100 and the maximum score is 100/100. This suggests that variance on the test is quite high, which leads to the second point. For most tests, the variance in the post-test is higher than the pre-test. This suggests that the measure is not as accurate. This in turn raises the third point, namely, that for a test out of 100, the confidence intervals are quite large, normally about 10 or more points. In short, the LLAMA test scores display a lot of variance, which may impact the accuracy of subsequent statistical tests.
5.1.5 Paired-samples t-tests (no groups)

A two-tailed paired-samples t-test was carried out for each LLAMA test to find any significant differences between pre- and post-test scores. Cohen’s $d$ was also calculated as a measure of effect size, that is, how big the difference was between pre- and post-test scores.

Table 5.4 reports the results from the paired-sample t-tests for all LLAMA tests.

Table 5.4 Results from paired-sample t-tests, 95% confidence intervals, and effect sizes (Cohen’s $d$) for each LLAMA test (B, D, E, F, and Total)

<table>
<thead>
<tr>
<th>LLAMA test</th>
<th>df</th>
<th>$t$</th>
<th>$p$</th>
<th>95% CI</th>
<th>Cohen’s $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>89</td>
<td>7.32</td>
<td>0.000</td>
<td>12.43-21.68</td>
<td>0.61</td>
</tr>
<tr>
<td>D</td>
<td>89</td>
<td>4.10</td>
<td>0.000</td>
<td>3.92-11.30</td>
<td>0.40</td>
</tr>
<tr>
<td>E</td>
<td>84</td>
<td>1.01</td>
<td>0.317</td>
<td>-2.59-7.88</td>
<td>0.11</td>
</tr>
<tr>
<td>F</td>
<td>89</td>
<td>2.07</td>
<td>0.041</td>
<td>0.30-13.99</td>
<td>0.21</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>5.57</td>
<td>0.000</td>
<td>23.10-48.73</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Results from the paired-samples t-tests indicate that post-test scores were significantly higher for the LLAMA B, $t(89) = 7.32, p =<.001, d = 0.61$; the LLAMA D, $t(89) = 4.1, p = <.001, d = 0.4$; the LLAMA F, $t(89) = 2.07, p = <.05, d = 0.21$; and for the LLAMA Total, $t(84) = 5.57, p <.001, d = 0.52$. Gain scores for the LLAMA E were not significant, $t(84) = 1.01, p = .317, d = 0.11$. Figures 5.7 and 5.8 show the box plots and figures 5.9 and 5.10 show the scatter plots for pre- and post-tests for each LLAMA test.
Figure 5.7 Boxplots for pre- and post-tests for LLAMA B, D, E, F
The size of the post-test score gain for each test was positive but varied in magnitude between tests. The 95% CI for the mean gain score for the LLAMA B was [12.43-21.68]. In the context of taking the test, this range indicates that the mean increase in post-test scores was at least 10 points up to about 20 points, out of a possible 100. Although the wide range in these scores makes it difficult to say exactly how much participants’ LLAMA B scores improved, it is important to note that LLAMA B scores are in increments of 5. The Cohen’s $d$ effect size calculated was 0.61, indicating that scores went up by three-fifths of one standard deviation, a small to medium effect size, according to Larson-Hall (2013). The 95% CI for the LLAMA

---

27 The current study follows Larson-Hall (2013) guidance for Cohen $d$ effect sizes of 0.40 as small, 0.70 as medium, and 1.0 and above as large.
D mean gain score was [3.92-11.3] and Cohen’s $d$ was 0.4, indicating a small effect size. The 95% CI for the LLAMA F mean gain score was [0.3-13.99] and Cohen’s $d$ was 0.21, indicating a very small effect size. The 95% CI for the LLAMA Total mean gain score was [-48.73–23.1] out of a possible 400 and with a Cohen’s $d$ of 0.52. This small to medium effect size indicates that mean gains overall for the LLAMA Total were in the magnitude of half a standard deviation. These variances and effect sizes are evident from the boxplots and the scatter plots, which include the 95% confidence intervals, as indicated by the shaded regions.

Figures 5.9 and 5.10 show LLAMA pre-test scores on the x-axis plotted against the post-test scores on the y-axis. Points have been jittered to show where more than one point is recorded for the same score as an indication of density of scores. The regression line is shown in red and the shaded region indicates the 95% confidence interval. The blue dashed-line shows a one-to-one relationship of pre- to post-test score to indicate an abstract of absolute stability in LLAMA test scores.
In general, each scatterplot shows a positive relationship between LLAMA pre- and post-test scores. In other words, for each LLAMA test high pre-test scores normally accompany high post-test scores, and vice versa. The regression lines tend to be flat with greater variance at the limits of the plot. This suggests that scores were most volatile for those who either scored very low or very high in the pre-test. Finally, post-test gain scores can be observed as points above the blue dashed-line, which is an abstraction of (absolute) stability in test scores. This clustering of scores above the diagonal from 0-100 gives an indication of post-test gain scores and is most noticeable for LLAMA B and LLAMA D. However, LLAMA F, by comparison, shows greater variance in post-test scores with more participants scoring lower in the
post-test. LLAMA E shows a ceiling effect with a substantial number of participants scoring 100% on the post-test.

**Figure 5.10 Scatter plot for LLAMA Total with the regression line (red) and its 95% confidence interval (blue shading)**

Overall, the results indicate that the change in mean scores from pre- to post-test was positive and significant for the LLAMA Total and for all individual LLAMA tests, except for the LLAMA E. The greatest gains were recorded for the LLAMA B and for the LLAMA Total, followed by the LLAMA D, with the smallest gains recorded for the LLAMA F. Effect sizes ranged from small to medium for the LLAMA B and the LLAMA Total to small for the LLAMA D and very small for the LLAMA F. In regards to the LLAMA E, no statistically significant gain was recorded, most likely due to a ceiling effect, with the pre-test mean being very high and a maximum score being less than one standard deviation from the mean \([M = 82.12, \ SD = 19.65]\). Importantly, however, despite the size of the gains, the LLAMA F
post-test scores were significantly higher than the pre-test scores, indicating that the training course could have been effective.

5.1.6 Paired-sample t-tests for low and high groups

Paired-samples t-tests were also carried out for all LLAMA tests with Low and High pre-test score groupings. A score was classified as Low if it was considered an average score or below as per the LLAMA test manual (Meara, 2005), and High if was considered above this average. The purpose of analysing Low and High pre-test score groups was to check if only those with initially low L2 aptitude were making the greater gains, as they had the most possibility to improve (Ganschow & Sparks, 1995; Politzer & Weiss, 1969; Sparks et al., 1998; cf. Yeni-Komshian, 1965). Numerical results are reported in tables 5.5 and 5.6, with t-test statistics, 95% confidence intervals, and effect sizes (Cohen’s d) for each group (Low and High) for each LLAMA test (B, D, E, F, and Total).

Table 5.5 Results from paired-sample t-tests for Low groups for all LLAMA tests

<table>
<thead>
<tr>
<th>LLAMA test</th>
<th>df</th>
<th>t</th>
<th>p</th>
<th>95% CI</th>
<th>Effect size (Cohen’s d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low B</td>
<td>37</td>
<td>5.02</td>
<td>0.000</td>
<td>13.34-31.40</td>
<td>0.64</td>
</tr>
<tr>
<td>Low D</td>
<td>56</td>
<td>5.55</td>
<td>0.000</td>
<td>7.96-16.96</td>
<td>0.60</td>
</tr>
<tr>
<td>Low E</td>
<td>4</td>
<td>0.95</td>
<td>0.397</td>
<td>-28.90-58.90</td>
<td>0.43</td>
</tr>
<tr>
<td>Low F</td>
<td>38</td>
<td>5.51</td>
<td>0.000</td>
<td>15.63-33.76</td>
<td>0.67</td>
</tr>
<tr>
<td>Low Total</td>
<td>20</td>
<td>7.21</td>
<td>0.000</td>
<td>53.73-97.51</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Results from paired-sample t-tests for High groups for all LLAMA tests show that gains were more prevalent in Low groups (LLAMA B, D, F, and Total) than High groups (LLAMA B and LLAMA Total only). Interestingly, medium effect sizes for the Low group individual tests were very similar (Cohen’s $d = 0.60-0.67$). Significant gains for both Low and High groups were only found for the LLAMA B and the LLAMA Total. Cohen’s $d$ effect sizes for LLAMA B Low and High were almost identical (0.64 and 0.63, respectively), indicating medium increases in post-test scores for all participants. The medium effect size for LLAMA Total Low ($d = 0.85$) was the largest of all post-test gains for all groups, although the LLAMA Total High had only a small gain in post-test score. This suggests that, on average, all participants increased their overall LLAMA test scores (LLAMA Total) from pre- to post-test and that these gains were greater for participants with Low pre-test scores. Gains for participants with Low pre-test scores appear to be approximately the same across the individual tests (excluding the LLAMA E).

The results from table 5.5. also indicate that variance in the 95% confidence intervals was greater for Low groups than High groups. In the case of LLAMA B Low and LLAMA F
Low, the range was approximately 18% of the total test score. While this volatility reflects smaller effect sizes, it also suggests the potential for substantial gains, as shown by the upper limit of the confidence interval.

Figures 5.11 and 5.12 show scatterplots of LLAMA pre- and post-test scores with the 95% confidence intervals shaded around the regression line.

Figure 5.11 Scatter plots for Low (triangle) and High (circle) groups for LLAMA B, D, E, F with the regression line and its 95% confidence interval (shading)
The scatter plots in figures 5.11 and 5.12 support the numerical data in table 5.5, while also showing the effects of the outliers in the data (see figures 5.5 and 5.6). The variance (and thus regression line) in LLAMA B are exaggerated by the outliers, which represent participants who scored 0 on the pre-test and then scored almost 100 in the post-test as well as those who scored just under 50 in the pre-test but 0 in the post-test. A similar situation can be seen with the outliers in LLAMA D who scored High on the pre-test but Low on the post-test. LLAMA E is particularly interesting, as the outlier who scored approximately 10 in the pre-test and 0 in the post-test has a strong influence, increasing the variance and pulling the regression line down into a steeper slope. The outliers in the LLAMA Total Low create a similar effect.

Interestingly, LLAMA F does not have any outliers, despite having greater variance overall. The distribution of scores in LLAMA F Low for participants who scored approximately 40 in the pre-test is also worth noting, with a considerable number scoring both lower and higher in
the post-test. Although the effect of the outliers appears substantial from the scatterplots, it
was beyond the scope of the current study to identify which factors caused their deviance
from the sample mean. Confounding factors such as motivation over the semester, testing
fatigue in the post-test, distractions in either pre- or post-test, etc. are almost impossible to
determine. Therefore, without any reasonable justification for excluding these outliers, it was
determined to retain them in the data set and remain mindful of their effects on the results,
which appear to both create and confound trends in the data (e.g. consider what LLAMA E
would look like without the extreme outlier at the bottom of the Low group).

In summary, the results from the paired-sample $t$-tests of Low vs High groups across all
LLAMA tests indicate that the gains identified at the whole group level were more prominent
for participants in the Low group for all tests, except the LLAMA E. In addition, the differing
effect sizes found between LLAMA tests at the whole group level almost disappeared at the
Low group level with LLAMA B, D, and F all showing medium effect sizes (Cohen’s $d$)
ranging between 0.60 to 0.67. However, the scatter plots show the influence of the outliers,
which must be taken into consideration when interpreting the results.

5.1.7 Pre- and post-test correlations

As a follow-up to the paired-samples $t$-tests, Pearson correlations ($r$) between pre- and
post-test scores were also calculated (see table 5.7). Correlations have been used in L2
aptitude studies to check for test-retest reliability (see Gliksman et al., 1979; Granena, 2013),
where strong correlations between pre- and post-test scores ($> 0.7$) are considered to indicate
that the measured abilities are stable over time.
Table 5.7 Pearson correlations between pre- and post-tests for all LLAMA tests

<table>
<thead>
<tr>
<th>LLAMA Test</th>
<th>r</th>
<th>95% CI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLAMA B</td>
<td>0.55</td>
<td>0.38-0.68</td>
<td>0.00</td>
</tr>
<tr>
<td>LLAMA D</td>
<td>0.45</td>
<td>0.26-0.61</td>
<td>0.00</td>
</tr>
<tr>
<td>LLAMA E</td>
<td>0.44</td>
<td>0.25-0.60</td>
<td>0.00</td>
</tr>
<tr>
<td>LLAMA F</td>
<td>0.27</td>
<td>0.06-0.46</td>
<td>0.01</td>
</tr>
<tr>
<td>LLAMA Total</td>
<td>0.49</td>
<td>0.30-0.63</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Correlations between all pre- and post-tests (n = 85) were significant; LLAMA B (r = 0.55, 95% CI [0.38-0.68], $R^2 = 0.3$), LLAMA D (r = 0.45, 95% CI [0.26-0.61], $R^2 = 0.2$), LLAMA E (r = 0.44, 95% CI [0.25-0.6], $R^2 = 0.19$), LLAMA F (r = 0.27, 95% CI [0.06-0.46], $R^2 = 0.07$), LLAMA Total (r= 0.49, 95% CI [0.3-0.63], $R^2 = 0.24$). These correlations are of medium effect size with a wide CI, which indicate that test-retest reliability is low (below 0.7). Interestingly, the correlation coefficients (r) progressively decrease from LLAMA B to F. This may, therefore, indicate test fatigue, although the fact remains that no test achieves an acceptable level of test-retest reliability. In other words, the test-retest reliabilities are not sufficiently acceptable to be able to claim that the pre- and post-test scores are not stable between pre- and post-test. This point is discussed in more detail in section 6.1.4.

### 5.1.8 Summary for research question 1

In summary, results from the paired-samples $t$-tests showed significant gains for all LLAMA tests, except for the LLAMA E. The size of gains ranged from small-medium (LLAMA B and LLAMA Total) to small (LLAMA D) to very small (LLAMA F). More detailed analysis
examining the gains based on groupings of low vs high pre-test scoring showed that participants with low pre-test scores made greater gains than those with high pre-test scores (LLAMA D, LLAMA F, and LLAMA Total), except for in relation to LLAMA B, where gains were similar for participants with low pre-test scores. Finally, Pearson correlations between pre- and post-test scores were significant, but relatively low \( r = 0.27-0.55 \) and with considerable variance. This indicates that test-retest reliability was under acceptable levels (0.7), that the post-test was not measuring the same ability or abilities as the pre-test, and that scores may be considered not stable between pre- and post-testing.

The second research question is dependent upon a significant gain in LLAMA F from pre- to post-test. Very small but significant gains were found for LLAMA F tests at the whole group level \( d = 0.21 \), but gains were substantially greater \( d = 0.67 \) for the Low group. This indicates that LLAMA F scores significantly improved over the 9 weeks of data collection, which included 6 weeks of language analytic abilities training. The second research question investigates whether the training course predicts these gains in LLAMA F post-test scores.

5.2 Research question 2: Do training factors explain more variance in post-test scores for language analytic abilities than L2 aptitude factors?

The second research question investigated the efficacy of the language analytic abilities training course in increasing LAAs, operationalised as the LLAMA F post-test scores. If the independent variables relating to the training course explain the variance in the LLAMA F post-test scores, this would suggest that the training was responsible for the gains in test scores. However, it is also possible that any or all of the LLAMA pre-test scores are also
important predictors of LLAMA F post-test scores, so these are also to be included in the model. If the training factors have higher relative importance in the model than the LLAMA pre-test scores, this would suggest that training was more predictive of increases than initial L2 aptitude levels and that training was at least partly responsible for the gains in language analytic abilities. To answer this question, a random forest regression model was fitted to discover how much of the variance in LLAMA F post-test scores can be explained by the variables of interest and also to determine which of the variables are the most predictive within the model.

In the current study, a random forest regression model was preferable over a linear model for three main reasons. First, the study included no control group, as it was impractical (see section 4.1 for discussion). This excluded the possibility of a more traditional pre-test, treatment-control, post-test design and, therefore, a more traditional ANOVA analysis to compare groups. Thus, a regression analysis looking at predictor variables to a response was preferable. If variables derived from the training course predict variance in LLAMA F post-test scores, this would indicate the training had an effect on language analytic abilities. Second, the study tracked real students engaged in a real-life learning context, not a highly controlled laboratory experiment. As a result, many participants did not attempt all tasks in the training and so no data were recorded for those unattempted tasks. When viewing these missing data points across all participants, no pattern of behaviour that would be informative to the analysis emerged. In other words, the training data were filled with missing values distributed unsystematically. A linear model would have excluded all data from participants with missing values, resulting in the loss of potentially useful data. Random forest models, on

---

28 While L2 learning experience was also seen as an important variable to investigate, there was not enough variation in the sample to pursue any analysis along these lines (see section 5.1.3 for details).

29 see Liaw & Wiener 2002 and chapter 4.6 for an explanation of random forests
the other hand, are much more tolerant of missing values as they aggregate predictions across a large number of classification trees fitted with bootstrapped samples, also allowing for missing values to be imputed (see Liaw & Wiener 2002 and section 4.6 for more details). Thirdly, due to the exploratory nature of the study, a large number of variables was of potential interest and importance. In order to retain explanatory power, linear models work best with a limited number of independent variables that minimises or avoids collinearity. Random forest models, by contrast, are based on tree models and are therefore not strongly influenced by collinearity. This makes random forest models particularly apt for exploring a lot of variables together. However, note that while random forest regression models are good at dealing with the problems mentioned above, the resulting model is an optimised one. Rather than offering a fully explanatory model, the results of random forest regression models are informative to the extent that they uncover predictors in the statistical sense.

A random forest regression model was fitted with the LLAMA F post-test as the response variable and with the independent variables as follows:

- pre-test scores for the LLAMA B, D, E, F, and Total
- participant approach to training in terms of total length and intensity of training
- percentage of training course completion
- participant accuracy scores over the training course
- participant time performance over the training course

### 5.2.1 Data preparation

Apart from the LLAMA test scores and the course completion, none of the independent variables of interest was able to be extracted or calculated directly from the raw data. The
reason for this is that the training course was designed first and foremost to be an educational and learning tool, not a pure statistical measure. As a result, raw data collected from the participants’ training had to first be prepared before being fitted to the final random forest regression model. The details of these preparations are detailed below.

5.2.1.1 Trimming times for training tasks and LLAMA tests

The time data collected from the training tasks and LLAMA pre- and post-tests contained a number of extreme outliers, with unusually large values above the median recorded. One possible explanation for this is that participants opened the training tasks and tests on their computers but were multitasking online and/or offline while also doing the training/LLAMA tests. The training/testing was then forgotten or not closed properly and so when the program window was eventually closed, a large completion time was recorded, despite the participant’s not having actively done anything for most of that time.

Instead of excluding these data, completion times were trimmed by the following process. First, the data were grouped according to task or test, for example, all times from the first task of week 1 were grouped together, the times from the first task of week 2, etc. Then within each group, robust weights were calculated and times that exceeded a limit of more than 99 minutes were then reassigned a value equal to the mean, that is essentially given a weight of zero. Maintaining the extreme outliers by reassigning them a value equal to the mean allows the sample size, and measures such as variance that are dependent on the sample size, to be maintained without distorting the mean of the core of the dataset. This process retained all the data without excluding or being confounded by extreme outliers from the dataset.
5.2.1.2 Initial aptitude level

Pre-test scores were taken directly from the raw dataset and included scores for the LLAMA B, LLAMA D, LLAMA E, LLAMA F, and for the LLAMA Total. Trimmed times (see above) were also included for the total time of the combined LLAMA pre-tests. Data from participants who had not completed all the LLAMA pre-tests and the LLAMA F post-test were excluded from the analysis. The final number of participants included in the sample was, therefore, n = 90.

5.2.1.3 Participants’ approach to training

To capture how participants chose to approach their training, a variable was created reflecting two dimensions. The first dimension captured the intensity, or frequency, with which the participants trained, operationalised as the average number of days between training sessions across the six weeks of the training course. In other words, if the participant normally massed all training tasks into one session once a week, the average number of days between sessions was seven; if the participant tended to space training sessions across all five weekdays, the average number of days between sessions was zero. The second dimension captured the total length of their training, measured in days from the first session of training to the last session. Therefore, if a participant started on the first day of the training period and finished on the last, the total duration of their training would have been 66; if they had done all the training in the last week, their total duration would have been seven. These two dimensions were then multiplied together to create a single measure of how each participant approached the training.
5.2.1.4 Training accuracy performance

To capture the performance of participants in terms of their accuracy scores across the training course, two levels of analysis were carried out. In the first level of analysis, scores from each training task were grouped according to task type. Six task types were used in the analysis. These task types appeared more than once during the training course and are as follows: pattern matching, problem solving, grammatical judgment task, multiple choice, processing input tasks, and Google Translate tasks (see section 4.4.3 for detailed explanations of each task type). Missing values were omitted and then each task type was fitted to a mixed-effects linear model. The fixed effect was the mean of the cohort of participants across each task. Each task type was checked to ensure that scores for all tasks within that task type differed significantly, before continuing the analysis. That is, the data were checked to ensure that participant performance across tasks was measurably different with enough variance for the model to be informative. This fixed effect can be thought of as the average performance across the training tasks, more accurately viewed as an average profile of performance for the group. The random effect, then, was the individual variance of participants from this profile of average performance, for each task type. This resulted in a continuous measure that indicated how well each individual participant performed on average across all tasks of each task type in relation to the mean performance of all participants. Thus, the first level of analysis produced a single measure of accuracy performance for each task type to be fitted to the random forest regression model.

In the second level of analysis, the results from the random-effects model underwent a principal components analysis (Crawley, 2013) to widen the scope of the analysis. The PCA resulted in two components with eigenvalues greater than 1, which are abstractions from the mixed-effects models of accuracy scores for each training task type. If the mixed-effects
models can be considered to capture the accuracy performance on the training of participants relative to the group, then the two principal components can be considered to reflect two aspects of this performance across all task types of the training. The first component is likely a reflection of ability, that is, how accurate in general a participant’s scores were for the course. The second component is less clear, but may possibly reflect the effort put into the test. However, this second more speculative component reflects another dimension in relation to which participants varied and so was considered potentially meaningful for the analysis. Consequently, the two principal components were considered as two summary variables identifying how participants varied in their accuracy scores across task types in the training course, and were fitted to the final regression models (see section 5.2.2).

5.2.1.5 Training time durations performance

To capture the performance of participants in terms of their time spent on each task of the training course, the same process of analysis as for the participant accuracy scores (see section 5.2.1.4) was carried out, with one exception.

Before grouping tasks by task type, missing values were summed for each participant. If the total of missing values for a participant exceeded 12, then the data from that participant were excluded from further analysis. Missing values in the remaining dataset were later imputed using an unsupervised imputation function adapted from the randomForest package in R (see Appendix C). When a participant had more than 12 missing values in their training tasks, then the imputed values were too distant from the actual values, that is, the imputed values were unreliable and would not add predictive power to the model (see section 5.2.2). The dataset of times spent on task was then structured according to task type, using the same six
task types as above. A mixed-effect linear model was then fitted to the data for each task type to create a mean profile of training time durations for the cohort, the same process as used in 5.2.1.4 to calculate task accuracy performance. The results of the fixed effects for each task type were checked to ensure that all tasks within that task type differed significantly, before continuing the analysis. That is, participants’ duration times across tasks needed to be measurably different to create enough variance to be informative for the final models. During this process, the differences between the Google Translate tasks were not found to be significant and so were excluded from further analysis. To reiterate, for a variable to be meaningful in a model, it needs to have significant variance. If the two tests had data for a task that were not significantly different from each other, then this insignificant or insufficient variance would render redundant the inclusion of such a task. Random effects for each task type were then calculated to represent the deviation of individual participants from the cohort mean. This resulted in a continuous variable for each task, representing participants’ performance on the training course tasks in terms of time taken, to be fitted to the random forest regression model.

The random effects calculated in the first level of analysis were then subjected to a principal components analysis (Crawley, 2013), again to widen the scope of the analysis. As per the training task scores, the PCA resulted in two components with eigenvalues greater than 1, which are abstractions from the mixed-effects models of accuracy scores for each training task type. If the mixed-effects models can be considered to capture the time durations performance on the training of participants relative to the group, then the two principal components can be considered to reflect two aspects of this performance across all task types of the training. The first component is likely a reflection of ability, that is, how quickly in general a participant was able to complete tasks over the whole course. The second
component is less clear, but may possibly reflect the effort put into the test. However, this second more speculative component reflects another dimension in relation to which participants varied and so was considered potentially meaningful for the analysis. Consequently, the two principal components were considered as two summary variables identifying how participants varied in their time durations across task types in the training course and were fitted to the final regression models (see section 5.2.2).

5.2.2 Random forest regression model

A random forest regression model was run with the LLAMA F post-test as the response variable and 23 potential predictor variables (see table C.7 in Appendix C for descriptive statistics on all variables in the regression models). Table 5.8 lists the 23 variables with a description of each term. The abbreviations used in the reporting of the analysis are also provided.

Table 5.8 The 23 variables included in the random forest regression model

<table>
<thead>
<tr>
<th>Variable type</th>
<th>Variable name</th>
<th>Variable abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLAMA Pre-test Score</td>
<td>LLAMA B</td>
<td>LLAMA_B</td>
<td>LLAMA B pre-test score</td>
</tr>
<tr>
<td></td>
<td>LLAMA D</td>
<td>LLAMA_D</td>
<td>LLAMA D pre-test score</td>
</tr>
<tr>
<td></td>
<td>LLAMA E</td>
<td>LLAMA_E</td>
<td>LLAMA E pre-test score</td>
</tr>
<tr>
<td></td>
<td>LLAMA F</td>
<td>LLAMA_F</td>
<td>LLAMA F pre-test score</td>
</tr>
<tr>
<td></td>
<td>LLAMA Total</td>
<td>LLAMA_Total</td>
<td>Aggregate score of LLAMA B, D, E, and F pre-test scores</td>
</tr>
<tr>
<td>LLAMA Pre-test Time</td>
<td>LLAMA Time</td>
<td>LLAMA_Time</td>
<td>Total time recorded to complete all LLAMA tests</td>
</tr>
<tr>
<td>Training Course Approach</td>
<td>Training Approach</td>
<td>Training Approach</td>
<td>How frequently and over how long a period of time participants trained (see section 5.2.1.3 for operationalisation)</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Training Course Completion</td>
<td>Training Completion</td>
<td>Training Completion</td>
<td>The proportion of all 26 training tasks participants completed</td>
</tr>
<tr>
<td>Training Task Accuracy</td>
<td>Pattern matching</td>
<td>TTA:Pattern</td>
<td>How well participants scored on average across all pattern matching tasks relative to the mean performance of all participants (see section 5.2.1.4)</td>
</tr>
<tr>
<td></td>
<td>Problem solving</td>
<td>TTA:ProblemSolving</td>
<td>How well participants scored on average across all problem solving tasks relative to the mean performance of all participants (see section 5.2.1.4)</td>
</tr>
<tr>
<td></td>
<td>Grammatical judgment task</td>
<td>TTA:GJT</td>
<td>How well participants scored on average across all grammaticality judgement tasks relative to the mean performance of all participants (see section 5.2.1.4)</td>
</tr>
<tr>
<td></td>
<td>Multiple choice</td>
<td>TTA:MCQ</td>
<td>How well participants scored on average across all multiple choice tasks relative to the mean performance of all participants (see section 5.2.1.4)</td>
</tr>
<tr>
<td></td>
<td>Processing input tasks</td>
<td>TTA:Processing</td>
<td>How well participants scored on average across all processing instruction tasks relative to the mean performance of all participants (see section 5.2.1.4)</td>
</tr>
<tr>
<td></td>
<td>Google Translate tasks</td>
<td>TTA:Translation</td>
<td>How well participants scored on average across all Google Translate tasks relative to the mean performance of all participants (see section 5.2.1.4)</td>
</tr>
<tr>
<td></td>
<td>Training Total Accuracy 1</td>
<td>TTA:Total 1</td>
<td>First dimension of relative training task accuracy for all task types (operationalised as first principal component of all)</td>
</tr>
<tr>
<td>Training Task Speed</td>
<td>Pattern matching</td>
<td>TTA:Pattern</td>
<td>How quickly on average participants completed all pattern matching tasks relative to the mean speeds of all participants (see section 5.2.1.5)</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------</td>
<td>------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Problem solving</td>
<td>TTS:ProblemSolving</td>
<td>How quickly on average participants completed all problem solving tasks relative to the mean speeds of all participants (see section 5.2.1.5)</td>
<td></td>
</tr>
<tr>
<td>Grammatical judgment task</td>
<td>TTS:GJT</td>
<td>How quickly on average participants completed all grammaticality judgement tasks relative to the mean speeds of all participants (see section 5.2.1.5)</td>
<td></td>
</tr>
<tr>
<td>Multiple choice</td>
<td>TTS:MCQ</td>
<td>How quickly on average participants completed all multiple choice tasks relative to the mean speeds of all participants (see section 5.2.1.5)</td>
<td></td>
</tr>
<tr>
<td>Processing input tasks</td>
<td>TTS:Processing</td>
<td>How quickly on average participants completed all processing instruction tasks relative to the mean speeds of all participants (see section 5.2.1.5)</td>
<td></td>
</tr>
<tr>
<td>Training Total Speed 1</td>
<td>TTS:Total 1</td>
<td>First dimension of relative training task speed for all task types (operationalised as first principal component of all training task speed measures, see section 5.2.1.5)</td>
<td></td>
</tr>
<tr>
<td>Training Total Accuracy 2</td>
<td>TTA:Total 2</td>
<td>Second dimension of relative training task accuracy for all task types (operationalised as second principal component of all training task accuracy measures, see section 5.2.1.4)</td>
<td></td>
</tr>
</tbody>
</table>
Training Total Speed 2

TTS:Total 2

Second dimension of relative training task speed for all task types (operationalised as second principal component of all training task speed measures, see section 5.2.1.5)

The random forest regression model was run with the following process and parameters (see section 4.6 for details of random forest models). Missing values were imputed to run an optimal model with 500 trees and randomly sampling 8 variables at each split. While imputing missing values does not make up for or gain data, it does allow for observed data to be retained in the dataset for analysis. Compared to discarding data which contained missing values, imputing these missing values increased the power of the analysis and reduced standard error by maintaining a larger sample size.

Table 5.9 shows the results of the random forest regression model. The number of trees indicates the number of bootstrapped samples used in the model. The number of variables at each split indicates the number of randomly sampled variables selected at each split. The mean of squared residuals is a measure of error in the model and the variance explained indicates how much variance in the model is predicted by the fitted variables.

Table 5.9 Results of random forest regression model predicting LLAMA F post-test scores

<table>
<thead>
<tr>
<th>Number of trees:</th>
<th>No. of variables tried at each split:</th>
<th>Mean of squared residuals:</th>
<th>% Var explained:</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>7</td>
<td>683.22</td>
<td>10.9</td>
</tr>
</tbody>
</table>
The results indicate that the model explains 10.9% of the variance in LLAMA F post-test scores, with a mean of squared residuals = 683.22, which in this case may be considered a substantial amount of error in the model.

The relative importance of the variables is shown in figure 5.13. Each variable is listed in order of contribution to the prediction of variance in the model. The higher up the list and the larger the value for the node purity (IncNodePurity), the more that variable contributes to the model’s prediction of the response variable, in this case, LLAMA F post-test scores.

The most important variable in the random forest regression model was Training Task Accuracy Total 1 (TTA:Total1), which most likely reflects participants’ ability on the training
tasks (see section 5.2.1.4). The LLAMA Total for the pre-test (LLAMA_Total) is fourth most important variable and the LLAMA F pre-test (LLAMA_F) is the ninth, nearly half as important as Training Task Accuracy Total 1.

Overall, the random forest model predicted only 10.9% of the variance in LLAMA F post-test scores. While this is only a small effect, the model nonetheless contains variables derived from the training course. Therefore, the model suggests that the training did have an effect on LLAMA F post-test scores, albeit a small one. This finding is supported by the results of the relative importance of the contribution of individual variables to the model. If L2 aptitude were stable, we could expect the LLAMA F pre-test score to be the most important predictor of LLAMA F post-test scores. The results show, however, that LLAMA F pre-test is the ninth most important variable, and, importantly, of the eight variables with more predictive power than the LLAMA F pre-test, five were derived from measures of the training course. Even more importantly, Training Task Accuracy Total 1, which can be considered a summary variable for the training course as a whole, was the most important predictor of LLAMA F post-test scores. This suggests the training had a positive effect on LLAMA F post-test scores.

5.2.2.1 Stepwise linear regression follow-up

As mentioned, the random forest regression model is an idealised and optimal model, representing the best possible analysis from the data given. As a follow-up, a stepwise linear regression model allows us to check that the most predictive variables are those given by the random forest model. To avoid inducing collinearity in the model, the LLAMA pre-test total and pre-test time duration as well as the two components for each of the training performance
variables, accuracy and speed, were omitted. The predictor variables of interest used in the stepwise linear regression model are as follows:

**Table 5.10 The 18 variables included in the stepwise linear regression model**

<table>
<thead>
<tr>
<th>Variable type</th>
<th>Variable name</th>
<th>Variable abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLAMA Pre-test Score</td>
<td>LLAMA B</td>
<td>LLAMA_B</td>
<td>LLAMA B pre-test score</td>
</tr>
<tr>
<td></td>
<td>LLAMA D</td>
<td>LLAMA_D</td>
<td>LLAMA D pre-test score</td>
</tr>
<tr>
<td></td>
<td>LLAMA E</td>
<td>LLAMA_E</td>
<td>LLAMA E pre-test score</td>
</tr>
<tr>
<td></td>
<td>LLAMA F</td>
<td>LLAMA_F</td>
<td>LLAMA F pre-test score</td>
</tr>
<tr>
<td>Training Course Approach</td>
<td>Training Approach</td>
<td>Training_Approach</td>
<td>How frequently and over how long a period of time participants trained</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(see section 5.2.1.3 for operationalisation)</td>
</tr>
<tr>
<td>Training Course Completion</td>
<td>Training Completion</td>
<td>Training_Completion</td>
<td>The proportion of all 26 training tasks participants completed</td>
</tr>
<tr>
<td>Training Task Accuracy</td>
<td>Pattern matching</td>
<td>TTA:Pattern</td>
<td>How well participants scored on average across all pattern matching tasks</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>relative to the mean performance of all participants (see section 5.2.1.4)</td>
</tr>
<tr>
<td></td>
<td>Problem solving</td>
<td>TTA:ProblemSolving</td>
<td>How well participants scored on average across all problem solving tasks</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>relative to the mean performance of all participants (see section 5.2.1.4)</td>
</tr>
<tr>
<td></td>
<td>Grammatical judgment</td>
<td>TTA:GJT</td>
<td>How well participants scored on average across all grammaticality judgement</td>
</tr>
<tr>
<td></td>
<td>task</td>
<td></td>
<td>tasks relative to the mean performance of all participants</td>
</tr>
<tr>
<td></td>
<td>Multiple choice</td>
<td>TTA:MCQ</td>
<td>How well participants scored on average across all multiple choice tasks</td>
</tr>
<tr>
<td>Training Task Speed</td>
<td>Pattern matching</td>
<td>TTS:Pattern</td>
<td>How quickly on average participants completed all pattern matching tasks relative to the mean speeds of all participants (see section 5.2.1.5)</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------</td>
<td>-------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>Problem solving</td>
<td>TTS:ProblemSolving</td>
<td>How quickly on average participants completed all problem solving tasks relative to the mean speeds of all participants (see section 5.2.1.5)</td>
</tr>
<tr>
<td></td>
<td>Grammatical judgment task</td>
<td>TTS:GJT</td>
<td>How quickly on average participants completed all grammaticality judgement tasks relative to the mean speeds of all participants (see section 5.2.1.5)</td>
</tr>
<tr>
<td></td>
<td>Multiple choice</td>
<td>TTS:MCQ</td>
<td>How quickly on average participants completed all multiple choice tasks relative to the mean speeds of all participants (see section 5.2.1.5)</td>
</tr>
<tr>
<td></td>
<td>Processing input tasks</td>
<td>TTS:Processing</td>
<td>How quickly on average participants completed all processing instruction tasks relative to the mean speeds of all participants (see section 5.2.1.5)</td>
</tr>
</tbody>
</table>
Table 5.11 shows the correlations of each predictor variable to the response variable of the LLAMA F post-test scores. These correlations show a simple one-to-one relationship between each predictor variable and the LLAMA F post-test scores and indicate which predictors are most likely to be important in the final linear model.

Of the 17 predictor variables, six were significantly correlated with LLAMA F post-test scores, and of those six, three were LLAMA (B, E, and F) pre-test scores. Interestingly, the strongest correlation was the LLAMA E pre-test \( (r = 0.36) \). Although the correlation is only small-medium, it is still higher than the LLAMA F pre-test (see section 6.2.3 for discussion).

The three training variables were Training Task Accuracy Pattern matching (TTA:Pattern) \( (r = 0.23) \), Training Task Accuracy Processing input tasks (TTA:Processing) \( (r = 0.30) \), and Training Task Accuracy Problem solving (TTA:ProblemSolving) \( (r = 0.24) \) task types, all of which showed small correlations with the LLAMA F pre-test.

Table 5.12 shows the results of the final stepwise regression model. The final model selected five of the 17 variables as the best combination for maximum prediction of the LLAMA F post-test scores (in order): LLAMA E pre-test, LLAMA F pre-test, Training Task Accuracy Processing input tasks (TTA:Processing), Training Task Speed Pattern matching (TTS:Pattern), and Training Task Speed Processing input tasks (TTS:Processing).
Table 5.11 Pearson correlations for predictor variables (listed) with response variable of LLAMA F post-test scores

<table>
<thead>
<tr>
<th>Variables</th>
<th>r</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLAMA_F_post</td>
<td>1.00</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>LLAMA_B</td>
<td>0.24</td>
<td>0.026</td>
</tr>
<tr>
<td>LLAMA_D</td>
<td>-0.19</td>
<td>0.078</td>
</tr>
<tr>
<td>LLAMA_E</td>
<td>0.36</td>
<td>0.001</td>
</tr>
<tr>
<td>LLAMA_F</td>
<td>0.27</td>
<td>0.009</td>
</tr>
<tr>
<td>Training_Approach</td>
<td>-0.11</td>
<td>0.289</td>
</tr>
<tr>
<td>Training_Completion</td>
<td>-0.04</td>
<td>0.728</td>
</tr>
<tr>
<td>TTA_Pattern</td>
<td>0.23</td>
<td>0.03</td>
</tr>
<tr>
<td>TTA_GJT</td>
<td>0.06</td>
<td>0.576</td>
</tr>
<tr>
<td>TTA_MCQ</td>
<td>0.10</td>
<td>0.339</td>
</tr>
<tr>
<td>TTA_Processing</td>
<td>0.30</td>
<td>0.004</td>
</tr>
<tr>
<td>TTA_ProblemSolving</td>
<td>0.24</td>
<td>0.025</td>
</tr>
<tr>
<td>TTA_Translation</td>
<td>0.14</td>
<td>0.173</td>
</tr>
<tr>
<td>TTS_Pattern</td>
<td>0.17</td>
<td>0.119</td>
</tr>
<tr>
<td>TTS_GJT</td>
<td>0.11</td>
<td>0.319</td>
</tr>
<tr>
<td>TTS_MCQ</td>
<td>0.15</td>
<td>0.165</td>
</tr>
<tr>
<td>TTS_Processing</td>
<td>0.04</td>
<td>0.709</td>
</tr>
<tr>
<td>TTS_ProblemSolving</td>
<td>0.07</td>
<td>0.503</td>
</tr>
</tbody>
</table>
### Table 5.12 Final model from stepwise regression predicting LLAMA F post-test scores

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SEB</th>
<th>β</th>
<th>Significance (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>18.1959</td>
<td>11.9159</td>
<td>1.527</td>
<td>0.13051</td>
</tr>
<tr>
<td>LLAMA_E</td>
<td>0.3283</td>
<td>0.1467</td>
<td>2.237</td>
<td>0.02793 *</td>
</tr>
<tr>
<td>LLAMA_F</td>
<td>0.2368</td>
<td>0.1032</td>
<td>2.295</td>
<td>0.02423 *</td>
</tr>
<tr>
<td>TTA:Processing</td>
<td>155.5734</td>
<td>54.5108</td>
<td>2.854</td>
<td>0.00544 **</td>
</tr>
<tr>
<td>TTS:Pattern</td>
<td>4.6424</td>
<td>2.0961</td>
<td>2.215</td>
<td>0.02948 *</td>
</tr>
<tr>
<td>TTS:Processing</td>
<td>-7.4747</td>
<td>2.9792</td>
<td>-2.509</td>
<td>0.01403 *</td>
</tr>
<tr>
<td>Multiple $R^2$</td>
<td></td>
<td></td>
<td></td>
<td>0.2693</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td></td>
<td></td>
<td></td>
<td>0.2258</td>
</tr>
<tr>
<td>$F(5,84)$</td>
<td></td>
<td></td>
<td></td>
<td>6.193</td>
</tr>
</tbody>
</table>

The stepwise linear regression model found that all final variables were statistically significant, with the final model accounting for approximately 23% of the variance (adjusted $R^2 = 0.23$). The strongest predictors of the LLAMA F post-test scores include (in order of predictive power) both aptitude (LLAMA E pre-test, LLAMA F pre-test) and training variables (Training Task Accuracy Processing input tasks, Training Task Speed Pattern matching, and Training Task Speed Processing input tasks).

In summary, the stepwise linear regression model found that a mix of LLAMA pre-test variables and training variables combined were the best model for predicting LLAMA F post-test scores. Like the random forest model, the amount of variance explained was small, but still significant. However, unlike the random forest model, the linear model found that L2 aptitude predictors were more important than training predictors. Importantly, the model
retained these training predictors, indicating that they were important predictors of LLAMA F post-test scores and that they contributed positively to the LLAMA F post-test scores.

5.2.3 Summary for research question 2

In summary, the random forest model explained 10.9% of the variance in LLAMA F post-test scores. Of the independent variables included in the model, Training Task Accuracy Total 1 (TTA:Total1) was the most important predictor, followed by the LLAMA Total pre-test scores. The LLAMA F pre-test score was not a strong predictor, relatively speaking, in ninth position of importance. This suggests that while the model had limited predictive power, it did account for some of the variance. In explaining that variance, training performance was the most important predictor, being almost twice as important as the LLAMA F pre-test. The stepwise linear regression did not completely support the results of the random forest model, with the L2 aptitude predictors being more significant predictors than the training predictors included in the final model. However, in both models, the vast majority of variance was not explained by either initial aptitude levels or training variables. Overall, the results do not conclusively show that the training course was successful in enhancing LLAMA F post-test scores, but they do suggest that the training did play a role, albeit a small one.

5.3 Summary of results for the research questions

Regarding the first research question, the results show that in general LLAMA L2 aptitude test scores were not stable and did significantly improve over the nine weeks of data collection. However, the gains were greater for learners with initially lower pre-test scores than for those with higher pre-test scores. Regarding the second research question, the results
suggest that the training course may have had an effect on the gains in LLAMA F post-test scores, but that this effect was small, and that LLAMA E and LLAMA F pre-test scores were also important predictors. These findings, their implications, and related issues are discussed in more depth in the following chapter.
CHAPTER 6: DISCUSSION

The current study investigated the purported stability and its correlate, untrainability, of L2 aptitude by attempting to answer the following main research question:

1. Is language learning aptitude stable or is it sensitive to experience?

This question was broken down into two more specific sub-questions, namely:

a. Using a pre-/post-test design, do post-test scores for L2 aptitude differ significantly from pre-test scores?

b. Do training variables explain more variance in post-test scores for language analytic abilities than L2 aptitude variables?

To answer these questions, a pre-test/intervention/post-test methodology was employed. L2 aptitude for research question 1.a was operationalised as the individual and cumulative scores for the LLAMA language aptitude tests (Meara 2005), while language analytic abilities were operationalised as scores for the LLAMA F test. All participants completed pre- and post-tests of the LLAMA tests and undertook a training course for language analytic abilities. Research question 1.a was investigated using paired-samples t-tests to analyse gain scores across all LLAMA tests. Research question 1.b used a random forest regression model to determine how much of the variance in LLAMA F post-test scores was explained by the data captured and to ascertain which variables were the most important. A follow-up analysis included a stepwise linear regression model to compare with the random forest regression model.

In general, results obtained in this study suggest that L2 aptitude increases with L2 learning experience and/or experience with L2 aptitude testing; that L2 aptitude gains are greater for
some factors than for others; that only 10.9% of gains in language analytic abilities can be explained by initial L2 aptitude levels and the variables of the training identified in chapter 5.2.2; and that training is an important predictor of gains in LAAs. However, it is unclear whether or not training is more predictive than initial L2 aptitude measures.

The following discussion will address each research question and their main findings. The discussion of research question 1 will address the gains in post-test scores, the size of those gains, differences in pre-test LLAMA F scores, and test-retest reliability. The discussion of research question 2 will address the predictive power of the random forest regression model, the relative importance of L2 aptitude and training variables, and the results of the linear regression. The discussion will then address the major implications of the findings for the theoretical construct of L2 aptitude in regards to the overarching question of stability in L2 aptitude, before concluding by identifying the most important challenges and limitations of the study.

6.1 Research question 1.a: Do post-test scores for L2 aptitude differ significantly from pre-test scores?

This section discusses the findings in relation to the first research question.

6.1.1 Pre- to post-test gains

Results from the paired-samples t-tests study support the view of L2 aptitude as dynamic and sensitive to experience (see 5.1.5). Significant gains were found for all LLAMA tests gains, except for the LLAMA E, most likely due to a ceiling effect in the pre-test results. These results suggest that participants improved in L2 aptitude abilities on work-sample tasks of
associative memory for vocabulary learning (LLAMA B), sound discrimination and memory for stretches of speech (LLAMA D), and language analysis for grammatical inference (LLAMA F) after only eight weeks of L2 Spanish learning with six weeks of training in language analysis. Gains were generally more evident for those participants whose initial pre-test scores were at or below average levels (according to the LLAMA test manual; Meara 2005), as opposed to those whose initial pre-test scores were above the average set by the manual (see 5.1.6). However, participants whose initial pre-test scores were above average on the LLAMA B and for the LLAMA Total also made significant gains.

These findings concur with those of previous studies. Sparks, Ganschow, and colleagues investigated the efficacy of a multisensory structured language (MSL) approach to L2 instruction for at-risk high school students (hereafter referred to as the MSL studies). They found that at-risk learners with MSL instruction and not-at-risk learners with traditional instruction achieved significant gains in MLAT test scores over a 12-month period (Ganschow & Sparks, 1995; Sparks & Ganschow, 1993; Sparks, Ganschow, Fluharty, & Little, 1995; Sparks, Ganschow, Pohlman, Skinner, & Artzer, 1992), with further gains also recorded over a two-year period for not-at-risk learners (Sparks, Ganschow, Artzer, & Patton 1997; Sparks, Artzer, Patton, Ganschow, Miller, Hordubay, & Walsh 1998). Gains were greater for at-risk learners, although the gap between at-risk and not-at-risk learners was never closed. While the findings for the current study are broadly in line with these MSL studies, there are three main differences. First, no participants with low pre-test scores were known to be at-risk learners, yet a divide was still found between this “low-level” group and those participants with high pre-test scores. This is unsurprising, as those with lower initial scores have more potential to improve than those with higher initial scores. Nevertheless, from the perspective of an assumption of stability, no such gains for either group should have
been evident. Second, the MSL studies were carried out over a much longer time, giving more opportunity for any instruction and learning to have an effect on L2 aptitude levels. However, gains in L2 aptitude in the current study were measured after just eight weeks, as compared with the 12 and 24 months of the MSL studies. This suggests that L2 aptitude may be extremely sensitive to any L2 learning experience (Thompson 2013). Third, the results of the MSL study could be questioned, as participants’ mean ages ranged from 14-16 years old and the MLAT is not a highly reliable measure for learners under grade 10 (approximately 15 y.o.) (Gliksman et al., 1979). The current study, by contrast, looked at adult learners with a mean age of 24 years old, thereby reducing the likelihood of any maturational effects on testing. Nevertheless, the results of the current study still broadly support the findings of the MSL studies, namely, that L2 aptitude levels increase with L2 learning experience.

Other studies have also shown suggestive gains from pre- to post-test. Yeni-Komshian (1965) found measurable increases in post-test scores\textsuperscript{30} for phonological abilities in adults after only two training sessions, although not all increases were statistically significant (as discussed in more detail below). In the only other L2 aptitude training study, Politzer and Weiss (1969) found measurable increases in post-test scores\textsuperscript{31} above test-retest gains for high school and adult L2 learners, suggesting the gains were above any test training effect (see 3.1.2.3). Interestingly, measurable gains were also found for a control group of high school students who completed the training but who were not undertaking any L2 studies at the time of L2 aptitude training, suggesting that the training had an effect on gain scores separate from any gains attributable to L2 learning experience. The authors speculated that L2 aptitude training was likely to be most beneficial prior to L2 learning, because L2 learning exerted a practice

\textsuperscript{30} Post-testing took place 1 week after training and 8 weeks after pre-testing. Instruments used included the MLAT II Phonetic script, MLAT III Spelling clues, and two trial tests that became part of the PLAB tests of phonological abilities.

\textsuperscript{31} Test-retest gains established for PLAB 3, 4, 5, and 6.
effect on L2 aptitude greater than explicit training. Unfortunately, the analysis did not include any comparison of paired-samples, only of treatment vs control, for which no statistical differences were found (discussed in more detail below). Demuth and Smith (1987) noted increases in L2 aptitude test scores\(^{32}\) (although no statistical tests were carried out) for 24 college students in the USA, identified as having L2 learning difficulties, who, prior to taking college L2 courses, undertook a novel series of (non-L2) courses that explained the L2 learning process and common problems L2 learners face. Sáfár and Kormos (2008) found that learners in a bilingual high school significantly improved compared with those of a control group in a non-bilingual school\(^{33}\). Despite similarly failing to include any paired-sample \(t\)-tests, Sáfár and Kormos concluded that some of the L2 aptitude tests were measuring trainable skills as opposed to stable abilities. This conclusion was based on studies that showed phonological sensitivity, which was measured in their L2 aptitude test battery, can be improved through direct instruction (Nijakowska 2008 as cited in Sáfár \& Kormos, 2008: 130). While Carroll (1981) has claimed to have empirical evidence that shows that gains achieved in the MLAT are temporary, the only published study found without significant gains between pre- and post-test was that of Moskovsky, Alshahrani, Ratcheva, and Paolini (2015)\(^{34}\). However, the lack of gains in this study might be attributable to the use of an unvalidated Arabic-language L2 aptitude test adapted from the MLAT.

Overall, the results of this study generally concord with those of the limited number of previous studies examining changes in L2 aptitude scores over time, the general conclusion of which is that L2 aptitude seems to increase with experience in L2 learning. These gains are more evident for learners with lower initial scores in L2 aptitude, which is perhaps

\(^{32}\) L2 aptitude instrument was the MLAT – presumably the Long form, although not stated in Demuth and Smith’s (1987) paper.

\(^{33}\) L2 aptitude instrument was the HUNLAT, a Hungarian adaptation of the MLAT.

\(^{34}\) L2 aptitude instrument was an Arabic adaptation of the MLAT.
unsurprising: this cohort has more potential to improve by virtue of having low scores in the first place. Gains also seem to be measurable in both the short and long term. Taken together, the results from pre- and post-testing L2 aptitude suggest that L2 aptitude is not stable, but rather is sensitive to L2 learning experience. More specifically, L2 aptitude test scores may be sensitive to experience, be it experience with previous L2 aptitude testing (see section 6.3.1 for discussion) or with L2 aptitude training treatments (see section 3.1.2.3) or with other areas of L2 experience, e.g. classroom learning. In direct contrast to the long-held assumption that L2 aptitude is stable and therefore untrainable (see section 3.1.1), the results of the current study show that L2 aptitude test scores change over time. Given that these tests are proxies for L2 aptitude abilities, the suggestion is that L2 aptitude itself is sensitive to experience.

6.1.2 Size of gains

The current study is the first to employ an analysis of effect sizes on the gains from pre- to post-test of L2 aptitude. Different effect sizes were found for each of the sub-tests, in order from largest to smallest: LLAMA B, a measure of associative memory and/or vocabulary learning; LLAMA D, a measure of sound discrimination and memory for speech patterns; and LLAMA F, a measure of language analytic abilities and grammatical inference (see section 5.1.5). LLAMA E post-tests were not significantly different from pre-test, so effect sizes are not discussed. On the understanding that L2 aptitude is what L2 aptitude tests measure (see section 2.1), when L2 aptitude test scores improve after L2 learning experience, the assumption is that L2 aptitude abilities are also improving. Interpreting the results from the perspective of Skehan’s (2002) model of second language acquisition and L2 aptitude, the
findings from this study suggest that L2 learning experience may effectively enhance L2 aptitude abilities (Ganschow & Sparks, 1995; Politzer & Weiss, 1969; Sáfár & Kormos, 2008; Sparks, Ganschow, Artzer, & Patton, 1997; Sparks, Ganschow, Fluharty, & Little, 1995). Skehan’s (2002) general argument is that learners draw upon different aspects of L2 aptitude depending on the stage of L2 acquisition at which they find themselves. For example, a beginner learner of L2 Spanish will initially rely upon phonemic coding and sound discrimination to acquire the Spanish sound system more than language analytic abilities to acquire the morphological or syntactic systems (Skehan 2002). In the current study, it is thus possible that in this beginning stage of L2 Spanish learning, participants’ L2 learning experience concentrated on memorising vocabulary (LLAMA B) and decoding the phonological system (LLAMA D) with less focus on analysing the grammatical system (LLAMA F). This would be consistent with the broadly communicative-inspired methodology followed in the university’s Spanish course (see section 4.2). The different effect sizes across the different LLAMA tests could thus be a reflection of the amount of ‘practice’ they had in these particular abilities, e.g. more memorising vocabulary and decoding phonology than analysing grammatical patterns. After eight to nine weeks of L2 Spanish learning, this training effect may have registered as measurable gains, with different gain sizes. In other words, L2 learning experience can train or enhance L2 aptitude (Politzer & Weiss 1969, Sáfár & Kormos 2008) (see discussion below for further elaboration). If this training effect is real, this finding also supports Skehan’s (2002) hypotheses that phonological and memory aptitudinal abilities are more implicated in the initial stages of L2 acquisition than language analytic abilities.

However, the results of the paired-samples t-tests for the low-high groupings challenge the idea of a differential training effect (see section 5.1.6). Comparable significant gains were
found for participants with below average pre-test scores with small-medium effect sizes for LLAMA B ($d = 0.64$), LLAMA D ($d = 0.60$), and LLAMA F ($d = 0.67$) and a medium effect size for LLAMA Total ($d = 0.85$). In contrast, significant gains for participants with above average pre-test scores were only found for LLAMA B ($d = 0.63$, comparable to LLAMA B Low) and LLAMA Total ($d = 0.37$). From this perspective, it would seem that only some individuals profit from L2 learning experience and that the gains are relatively similar across the various tests for those that do improve. Interestingly, however, the LLAMA Total Low group only contained 13% of participants, while the LLAMA B, D, and F Low groups contained anywhere between approximately 40-60% of participants. This result suggests that most individuals had weaknesses in only one or two tests that were balanced by strengths in others. Thus, the gains for the LLAMA Total High might be more accurately interpreted as an approximate average of gains possible for participants experiencing this particular L2 learning situation. Nevertheless, the similarity of gains across the low sub-groupings of participants casts doubt upon any differential training effect found when analysing gains only at the cohort level. More research that includes effect sizes for objectively quantifying gains from pre- to post-test is needed to determine whether training effects from the L2 learning experience are differential across the various L2 aptitude abilities.

6.1.3 LLAMA F differences for L1-L2 level

Results from the analysis of variance of background variables also found that both L1 English and non-L1-English participants with extensive previous L2 learning experience – proficient in at least one L2 plus experience in learning another language – scored significantly higher in the LLAMA F than those with less experience (see section 5.1.3). However, no significant
differences emerged for any other LLAMA test (B, D, E) or background variable (age, L1). This finding is broadly in line with other studies with cross-sectional analyses (see section 3.1.2.2), which found that the number of languages learned correlated with L2 aptitude scores (Eisenstein, 1980; Grigorenko et al., 2000; Planchon & Ellis, 2014; Thompson, 2009), that learners with L2 learning experience scored higher on L2 aptitude tests than monolinguals (Carroll, 1967; Natelson & Allen, 1977), and that the number of languages learned could be a better predictor of L2 achievement than the MLAT (Brecht et al., 1993). Taken together, the findings from this and previous studies on changes in L2 aptitude scores, strongly question the validity of the assumption of stability in L2 aptitude and L2 aptitude testing. However, it remains unclear why differences were only found for the LLAMA F. One possible explanation may relate to Granena’s (2013) characterisation of the LLAMA F as a measure of an L2 aptitude for explicit L2 learning. The LLAMA F test task is designed with a learning phase of 5 minutes followed by an untimed testing phase. The task is designed such that the learning of the grammatical patterns can be done in a variety of ways. Therefore, strategic approaches to the task might be very effective. As discussed in 3.1.3.4, different types of strategies can be taught/learned and can distinguish multilingual L2 learners from less experienced L2 learners (McLaughlin, 1995). On this basis, the differences that emerged in relation to LLAMA F might be attributable to its accommodation of a variety of strategic approaches to L2 learning that are developed with extensive L2 learning experience. Conversely, these differences did not emerge in relation to the tasks of any of the other tests, which did not accommodate such a variety of strategies, and which thereby rendered L2 learning experience less beneficial to achieving higher scores. More research in this area is needed to understand what effect previous L2 learning experience has on individuals taking L2 aptitude tests.
6.1.4 Test-retest reliability

Correlation tests of pre- and post-test scores are normally used to check the reliability of a measure over time. Test-retest reliability is thus a taken as evidence that the measured abilities are stable over time (e.g. Granena, 2013; Gliksman et al., 1979). A follow-up analysis looking at the correlations between pre- and post-test scores in the current study found that all tests were significantly correlated, although correlations were relatively low and displayed wide variance, with none being higher than $r = 0.55$ (LLAMA B) (see section 5.1.7). This suggests that the test scores were not stable over time. In her exploratory validation of the LLAMA test, however, Granena (2013) found statistically significant correlations between pre- and post-test scores over a two-year period ($n=20$, mean age = 25 y.o.) between 0.53-0.64, which she considered as minimally sufficient to indicate stability.

Comparing each LLAMA test from the current study with Granena’s (2013), only the LLAMA B had similar results, $r = 0.53$ (Granena 2013) and $r = 0.55$ (current study). Similarly, Gliksman, Gardner, and Smythe (1979) found MLAT test-retest correlations for high school students in Grade 10 and 11 between 0.60-0.72 over a one year period. Plainly, while not optimal, the results of both previous studies are minimally acceptable for test-retest reliability. In the present study, however, the results clearly show that scores were not stable over a much shorter period of time than previous studies (LLAMA B $r = 0.55$, LLAMA D $r = 0.45$, LLAMA E $r = 0.44$, LLAMA F $r = 0.27$, LLAMA Total $r = 0.49$). This result is able to be interpreted in two ways. The first possible interpretation is that the LLAMA test is not an adequately reliable measure (see section 3.1.2). However, while the LLAMA test remains unvalidated, Granena’s (2013) results did meet minimum standards for test-retest reliability, which is impressive, considering that the study was conducted over a two-year period. Consequently, Granena’s results suggest that the LLAMA is reliable in general. The second
possible interpretation is that the participants approached the post-test in a measurably different way. As Snow and Lohman (1984) explain, in a dynamically cognitive system, individuals can profit from experience and adapt to the demands of the learning task. For example, from their L2 Spanish learning experience, from their training in language analysis, or from a combination of these activities, participants may have gained insight into the demands of the LLAMA test tasks and adapted their approaches in the post-test. Thus, the low correlations may not reflect an unreliable measure, but rather that the post-test scores suggest that participants completed the test in a qualitatively different manner to the pre-test (see section 3.1.3.3 for further discussion). More research is needed to clarify these issues of test-taker behaviour on L2 aptitude tests and its impacts on test-retest reliability measures.

6.1.5 Summary

Overall, an examination of whether L2 aptitude post-test scores differ significantly from pre-test scores, has revealed that on practically all measures, L2 aptitude scores increase with L2 learning experience and explicit training. These increases were measurable after just eight weeks of L2 learning with adult learners, with differential gains between the different measures of L2 aptitude at the cohort level. Gains were greater for those whose pre-test scores were below average and, interestingly, the size of the gains across the different LLAMA tests for this sub-group were comparable. Thus, in contrast to the cohort-level analysis, the sub-group analysis suggests that the size of the gains was relatively similar for those who did make gains. These longitudinal results were partially supported by the finding that pre-test scores for the LLAMA F were significantly higher for those participants with extensive prior L2 learning experience, although it is unclear why differences for the other tests were not found. Finally, correlations between pre- and post-tests were significant but
small, suggesting that the LLAMA tests are either unreliable measures or that the post-test evidences the use of a qualitatively different approach in the post-test compared to the pre-test. In short, the results for the first research question generally suggest that L2 aptitude is in fact sensitive to L2 learning experience. This entails that L2 aptitude is also trainable and strongly questions the theoretical assumption of stability.

6.2 Research questions 1.b: Do training factors explain more variance in post-test scores for language analytic abilities than L2 aptitude factors?

This section discusses the main findings in relation to the second research question.

6.2.1 Predictive power of random forest model

Research question 1.b looked at the possible efficacy of the experimental training course in language analytic abilities. The results of the random forest regression model explained 10.9% of the variance in the LLAMA F post-test scores, including predictors of initial L2 aptitude, amount of training completed, approach to training, and performance on training in terms of speed and accuracy (see section 5.2.2). While the random forest model is optimised, two facts suggest that the training was at least minimally effective. Although the amount is small, some of the variance in LLAMA F post-test scores was predicted by the training variables. Of the 23 variables in the random forest model (see table 5.7), Training Task accuracy 1 – considered a summary measure of ability for the whole training – was the most important predictor. In the stepwise linear regression analysis, the final model retained three variables from the training: Training Task Accuracy Processing input tasks, Training Task
Speed Pattern matching, and Training Task Speed Processing input tasks (see section 5.2.2.1). These results suggest that the training did have some effect on LLAMA F test scores, otherwise the training variables would not have been meaningful predictors. More research is required to determine whether this trainability relates only to a certain aspect of language analytic abilities or to a certain stage of development. Nevertheless, the results from the random forest model indicate that gains in LLAMA F test scores were partially accounted for by training variables, thus suggesting that language analytic abilities are trainable, to some extent.

The two previous L2 aptitude training studies also support a view of L2 aptitude as trainable. Politzer and Weiss (1969) found that high school and adult participants undertaking a custom-designed L2 aptitude training increased their MLAT and PLAB scores above re-test levels, despite having negative attitudes toward the training and a lack of strong motivation. Gains were also measured for a control group that had not been undertaking any L2 learning at the time of the training, suggesting that the training was effective in increasing L2 aptitude test scores. Among participants in the experimental treatment, those with lower pre-test scores made greater gains than those with higher pre-test scores, which is consistent with the findings of the current study (see section 5.1.6). Yeni-Komshian (1965) found similar trends in her training study of high school L2 learners undertaking phonological training. Again, gains were measurable for experimental and control groups across many of the L2 aptitude phonological abilities, although only some were significant. Experimental and control groups did not differ significantly in the size of their gain scores. Interestingly, Yeni-Komshian found that the greatest gains were among participants with high pre-test scores of about 80%, contradicting the findings of Politzer and Weiss and those of the current study. While both studies were exploratory in nature and stated the need for further research to substantiate their
findings, both concluded that there did indeed seem to be some benefit to be derived from undertaking L2 aptitude training prior to starting L2 learning. Despite the extremely small number of L2 aptitude training studies, all three – Politzer and Weiss (1969), Yeni-Komshian (1965), and the current study – have shown suggestive evidence that L2 aptitude in general, as well as phonological abilities and language analytic abilities specifically, are trainable.

6.2.2 Relative importance of predictors

The random forest model also provided an ordered list of the most important variables for predicting the response variable of the LLAMA F post-test scores. In total, 23 variables (see section 5.2.2 and table 5.7 for a complete list) were included, among which were variables that covered initial L2 aptitude levels, approach to the training course, how much of the training course was completed, and the training performance in terms of accuracy and speed across six training task types. Of the eight most important variables, five related to performance on the training, while only three related to L2 aptitude. Training variables that were most important were (in order of importance): Training Task Accuracy Total 1, Training Task Accuracy Processing input tasks, Training Task Accuracy Multiple Choice, and Training Task Accuracy Problem Solving. Two reasons which possibly account for the relative importance of these particular test tasks are as follows. First, each of these task types may have involved the use of strategies, which was discussed above as being possibly an element to higher achievement in the LLAMA F test task (see section 6.1.3). For example, the Processing input tasks involved learning and practicing strategies for overcoming common L2 processing difficulties (see section 3.2.4); Multiple choice tasks are presumably open to test-wiseness strategies of eliminating less likely answers to narrow the range of possible answers (see section 2.1.2); and Problem solving tasks invite the use of metacognitive skills
such as planning and strategy selection (see section 3.2.4 for discussion). Second, and correlatively, perhaps it is precisely the range of abilities implicated in these task types that makes them important predictors of LLAMA F post-test performance. The L2 aptitude variables that were most important included (in order of importance): LLAMA Total pre-test scores, LLAMA E pre-test scores, and the LLAMA Time for the pre-test. The LLAMA F pre-test is notable here by its absence. If L2 aptitude were indeed stable, we would assume that the LLAMA F pre-test scores would be the strongest predictor of LLAMA F post-test scores. The fact that the LLAMA F pre-test scores were less important than a range of training variables suggests that the training was responsible for some of the gains in the LLAMA F post-test scores. To confirm or disconfirm these findings, we now discuss the results from the linear model.

6.2.3 Linear model in comparison

The results from the linear model were included as a check on the random forest model results, given that the latter statistical model is relatively infrequently used in applied linguistics. However, the results contradicted the random forest to a certain extent, with L2 aptitude variables being more predictive than the training variables. With the linear model, the L2 aptitude variables did include the LLAMA F pre-test, but this last was still less predictive than the LLAMA E pre-test. Once again, the training variables included the relative performance for accuracy for structured processing input, but also included the two variables of relative speed performance for structured processing input and for pattern matching. The inclusion of these two speed variables is interesting in that the relationship between time durations and LLAMA F post-test scores is inverse, i.e., the lower the time duration measure, the higher the LLAMA F post-test score. In other words, the faster
participants performed on the structured processing input tasks, the higher they scored on the LLAMA F post-test. By contrast, those participants who spent more time on the pattern matching tasks, i.e., were slower, scored higher on the LLAMA F post-test. This suggests that answering quickly on the structured processing input tasks indicated an ability to quickly analyse deductively from a rule, whereas answering slowly on the pattern matching tasks may suggest a more methodical and exhaustive type of inductive language analysis, both skills resulting in better performance on the LLAMA F post-test. More research is needed to confirm this hypothesis. As mentioned, the overall results of the linear model regression challenged those of the random forest model. However, several caveats should be borne in mind when considering the nature and implications of this challenge. The limitations discussed below severely restrict the explanatory power of the results of the linear regression model and should only be taken as suggestive of relationships that warrant further research.

6.2.4 Summary

Overall, in attempting to determine whether training factors explain more variance in post-test scores for language analytic abilities than L2 aptitude factors, the results are inconclusive. While the random forest model found that Training Task Accuracy Total 1 was the most important predictor of the LLAMA F post-test scores, the linear model found that LLAMA E and LLAMA F pre-test scores were the best predictors. Importantly, however, in both models initial L2 aptitude variables and training variables were important predictors. Thus, while the results were unclear as to their relative importance, training variables were still important and significant predictors of LLAMA F post-test scores, suggesting that language analytic abilities are amenable to training. However, it must be recalled that the explanatory power of the statistical tests run is limited, as these last were optimised and
focussed on finding the best predictive variables. It would seem, therefore, that the most important finding is that while initial L2 aptitude and training performance are important predictors, these factors still account for only a small proportion of the variance in LLAMA F post-test scores. As mentioned, one possible explanation for this might be that participants gained an understanding of the language learning process through experience or developed more effective L2 learning strategies (see section 3.1.3.4), although there is very little evidence to support such a view. Thus, the underlying causes of the significant gains from pre- to post-test scores on the LLAMA F remain unknown and constitute an important area for future research.

6.3 Implications of the study

The results of the current study show the need to expand Carroll’s notion of L2 aptitude. Methodological objections, such as sample characteristics (e.g. homogeneity and matching) and a lack of comparison between pre- and post-test correlations with L2 achievement criterion (Carroll, 1962, 1974, 1981, also see section 3.1.1), might be raised to refute the current study’s findings. However, for the purposes of SLA studies, Carroll’s construal of L2 aptitude as consisting solely of stable first-order abilities appears to be at odds with the reality of most L2 aptitude tests (see section 2.1.1). Rather than accurately isolating and measuring aptitude-as-first-order-abilities, these work-sample tests allow for higher-order thinking, such as employing strategic approaches which can be adapted throughout the tests. For the purposes of SLA studies, Carroll’s restrictive (and unsubstantiated) conception of L2 aptitude needs to be reconstrued and expanded to explicitly include the higher-order abilities whose use is implicitly elicited by most L2 aptitude tests. Importantly, these higher-order abilities in
L2 aptitude have been shown in the current study to be amenable to training and thereby improvement. On this basis, an expanded construal of L2 aptitude seems warranted, one which would serve SLA studies more usefully than the current view. The current thesis has identified some of the basic concepts that might underpin such an expanded construal by proposing cognitive-based models of L2 aptitude that are hierarchical, in which L2 learners are seen as adaptive agents responding to feedback obtained through their L2 learning experience, and that higher-order skills and knowledge of L2 aptitude, namely metacognitive strategies and knowledge of L2 processes, can be taught.

An expanded construal of L2 aptitude that includes higher-order abilities and that assumes that L2 aptitude can improve with experience holds important implications for L2 aptitude measurement. Carroll’s construal of L2 aptitude as stable first-order abilities implies the use of a one-off static testing of an individual, seemingly ignoring experiences from L1 and L2 learning. This view of L2 aptitude is at odds with MLAT-style L2 aptitude tests that allow for the use of strategies and in which knowledge of L2 processes may give test-takers an advantage. If these higher-order abilities of L2 aptitude can be acquired by experience and/or instruction, then perhaps L2 aptitude is better construed as the potential to learn, a view adopted in dynamic testing. Dynamic testing entails testing on more than one occasion and providing examinees with feedback and interaction from which to profit and learn (Grigorenko & Sternberg 1998). In other words, dynamic testing is interested in “whether and how the subject will change if an opportunity is provided” (Grigorenko & Sternberg 1998:76). Plainly, the construal of L2 aptitude as sensitive to experience and amenable to training more closely aligns with the perspective of dynamic testing than with the Carrollian view of static testing.
Assuming that the results of the current study are reliable, two considerations for L2 aptitude research present themselves. The first is that the gains measured in the post-test were a result of the experience in test-taking, that is, that the pre-test had a training effect on post-test performance. The second is that the gains in the post-test were a result of experience in L2 (Spanish) learning (including any effect of instruction and training in LAAs), that is, the learning of Spanish had a training effect on post-test performance. Each of these implications will be discussed below in more detail with a particular focus on L2 aptitude testing.

6.3.1 Test training effect

The validity of L2 aptitude tests seems to rely heavily on the novelty of the work-sample tasks presented to test-takers. Carroll (1962) cites studies where L2 aptitude test and criterion correlations were low, because participants had previous L2 learning experience. Assuming that the work-sample tasks are valid representations of L2 learning tasks, then L2 aptitude measures are highly sensitive to knowledge of how to solve such problem sets. Any experience with this type of task, therefore, would offer the test-taker a learning opportunity to improve their ability with these kinds of problems. This is what Gliksman, Gardner, and Smythe (1979) suggest from their investigation of the test-retest reliability of the MLAT for high school students. Their findings suggest that familiarity with the test itself was a likely explanation for the low correlations between test and retest scores taken a year apart. These findings are in line with the current study and would suggest that the gains on the post-tests are a result of the familiarity with the pre-test. Low test-retest reliabilities may also be a result of maturational effects (e.g. Harley & Hart 1997). However, while some pre-/post-test findings are susceptible to maturational effects of the teenage participants (see Ganschow & Sparks, 1995; Gliksman et al., 1979; Sáfár & Kormos, 2008; Sparks & Ganschow, 1993;
Sparks, Ganschow, Artzer, & Patton, 1997; Sparks, Artzer, Patton, Ganschow, Miller, Hordubay, & Walsh, 1998; Sparks, Ganschow, Fluharty, & Little, 1995; Sparks, Ganschow, Pohlman, Skinner, & Artzer, 1992), the adults of the current study do not fit this profile nor those of other pre-/post-test studies (Demuth & Smith, 1987; Politzer & Weiss, 1969; Yeni-Komshian, 1965). Nevertheless, there is a logical problem here. L2 aptitude tests are taken as proxies of potential L2 achievement based on the assumption that the work-sample tasks are fair representations of real L2 learning problems. Following this assumption then, if familiarity, or experience, has a training effect on test performance, then it follows that L2 aptitude is a skill that is sensitive to experience and can be trained. Thus, we find Carroll’s (1981) model of L2 aptitude as stable too limited and needs to be expanded to account for higher-order abilities of L2 aptitude.

6.3.2 L2 experience as training effect

On the other hand, dynamic testing would take a fundamentally different view of what L2 aptitude is. Rather than a single test constituting a measure of potential L2 achievement, dynamic testing takes the view that any changes in scores between testing are what constitute a learning ability (Grigorenko & Sternberg, 1998). Any change is considered a learner response to feedback, typically in the form of instruction, but seemingly not precluding learner adaptation based on task performance. In other words, receiving their marks from the pre-test – feedback on the effectiveness of their approach – may have triggered a change in participants, leading to better or worse performance in subsequent testing. Thus, L2 aptitude is not what you know about the various types of L2 learning problems, but rather how you respond to the problems and, in particular, how you profit from feedback and/or instruction.
Sternberg (2002) cites a study in support of this perspective in which young children in Tanzania were pre- and post-tested with Western-style tests of general aptitude that were foreign to their educational culture. Between pre- and post-testing, the children were given 15 minutes of instruction in related skills that would benefit them on the post-test. Significant gains were found between pre- and post-test scores despite a short time period between testing. More interestingly, correlations between pre- and post-test scores were low, roughly $r = .3$, demonstrating how unreliable a static measure of general aptitude was in this context, which is consistent with the results of the current study (see sections 5.1.7 and 6.1.4). Because the style of test was so unfamiliar to the children, any instruction had obvious effects. This type of unfamiliarity is analogous to naive language learners taking an L2 aptitude test for the first time. Thus, any feedback on the test is likely to be extremely beneficial and significant gains between pre- and post-test are thus highly possible. This description would seem to fit the current study, particularly in light of the fact that most participants seemed to have evinced a weakness in either the LLAMA B, D, and F. Thus, participants may have been able to improve weaker pre-test scores through responding to the feedback they received. Thus, a dynamic testing perspective may offer a sufficient explanation of gains in post-testing scores of L2 aptitude tests.

In short, the results from this study highlight a disconnect between L2 aptitude theory and testing. L2 aptitude construed as stable first-order abilities entails static one-off testing, yet there is little evidence to support the view of current L2 aptitude tests as isolating measures of pure first-order abilities. In fact, this thesis has argued for L2 aptitude as needing to be construed as more than just stable first-order abilities. In contrast to Carroll’s model, L2
aptitude construed as including higher-order abilities can be measured with dynamic and/or reiterative testing regimes that track changes in L2 aptitude over time.

Following the argument for dynamic testing above, it is also possible that the gains in pre-test were not driven by experience in test-taking alone, if at all. Instead, it is possible that the experience of learning L2 Spanish operated as a form of instruction bringing about the measurable gains in L2 aptitude. Another potential form of instruction includes the training course in language analytic abilities.

Apart from any potential effects of instruction on post-test scores, feedback to the learner may also have been important. After completing the pre-tests, participants were supplied with their results and explanations of how the tests related to real L2 learning. This feedback may have also triggered changes in participants. For example, upon learning of any weaknesses in their L2 aptitudinal abilities, participants may have been sensitive to their challenges in learning L2 Spanish and responded in some way. This would entail some level of adaptation by the participant, resulting in a measurable change in their approach to L2 learning and, thus, L2 aptitude. While such speculation is feasible according to a dynamic testing perspective of L2 aptitude, more research is needed to investigate these possibilities and alternative perspectives of L2 aptitude.

Although the above discussion regarding the current study is purely speculative, there is some support for construing L2 aptitude as a dynamic construct. Most obvious is the CANAL-FT (Grigorenko et al., 2000), which is a dynamic test of L2 aptitude test. The CANAL-FT can be considered as valid a measure of L2 aptitude as the MLAT, with highly significant strong correlations between the two tests (Grigorenko et al., 2000). In a very different study, Krug, Shafer, Dardick, Magalis, and Parenté (2002) found that performance on an oral
paired-associates vocabulary test iterated 10 times in a row was an effective classifier of L2 learning ability. The test was extremely simple, consisting of 10 L1 items with L2 translations, but was able to accurately classify L2 learners by ability. As has already been discussed, a body of evidence also exists showing measurable gains between pre- and post-tests of various L2 aptitude measures (Demuth & Smith, 1987; Ganschow, 1993; Ganschow & Sparks, 1995; Sáfár & Kormos, 2008; Sparks et. al., 1992; Sparks et. al., 1997) that would also support a dynamic view of L2 aptitude as sensitive to experience and training. Finally, the fact that 90% of the variance in post-test scores was not accounted for by the random-forest regression model is also an important finding with interesting implications. Obviously, the unaccounted variance in the LLAMA F post-test scores indicates that an unanticipated variable or unanticipated variables were the drivers behind the gains in the scores. Adaptation and training effects from the L2 learning experience have already been discussed, but it is also possible that a more individual factor was responsible for gains, such as engagement in L2 learning. For example, participants who were stimulated by their L2 Spanish learning may have engaged in the process to a level that brought about improvements in their L2 aptitude and particularly their language analytic abilities. This engagement may have resulted in participants undertaking more learning, increasing the likelihood of encountering challenges to their learning strategies that may have necessitated some level of adaptation in order to progress. This adaptation may have been responsible for the improvement in language analytic abilities. The identification and examination of individual factors contributing to gains in L2 aptitude test scores, seems a rich and important area of future research (see Rogers et al., 2016; Rogers et al., 2017).
6.4 Challenges and limitations of study

Unsurprisingly, perhaps, the current study faced many challenges, of which two are particularly noteworthy. First, the nature of the data collected made using a standard linear model extremely challenging. With so many variables potentially being important predictors of LLAMA F post-test scores, and with very few clues from previous research or theory to exclude some of these variables, an alternative random forest regression model, relatively unknown in applied linguistics, had to be used. The use of the random forest regression allowed for greater insight into the data and provided a more optimised model of the data than the stepwise linear regression. Second, the administration of the LLAMA test is not optimal and had to be adapted to be delivered online. Administering the test in this way meant that testing could not be strictly standardized, relying on participants following the instructions accurately without supervision. These administrative issues were unavoidable, however, in order to collect data from a sufficiently large sample for statistical analysis. The opacity of the LLAMA test’s marking system also presented problems. Specifically, the LLAMA test programs automatically generate a final mark for test-takers, but do not include specifics for which questions were answered correctly and incorrectly, or a total of how many questions were correct and incorrect. This resulted in the marks including less detail of participants’ actual performance on the test, which may have proved useful in the statistical analysis. For example, no internal reliability measures could be generated for the LLAMA test without these data.

As previously discussed (see section 4.7), the study also has some distinct restrictions that limit the interpretation of results and their generalizability. First and foremost, the study was

---

35 It should be noted that this study took place before Rogers, Meara, Barnett-Legh, Curry, and Davie (2017) was published with information about the marking system employed by the LLAMA tests.
essentially a field experiment of a real L2 learning situation. While this increases the ecological validity of the study, it comes at the expense of experimental validity. Second, collecting data from participants engaged in a real life learning situation also makes for messy data with a lot of missing values and some level of self-selection, as participants were free to desist if they chose. Third, the LLAMA test is still not completely validated as a measure of L2 aptitude (Rogers et al., 2017) and a more detailed and accurate record of marking in the LLAMA tests would have added useful information to the analysis, i.e. a record of how many questions were answered right and how many wrong as opposed to a summary score. Fourth, sampling was one of convenience and was susceptible to the normal problems of self-selection. Lastly, it is also worth noting that the current study encountered various complications. Of these, the most important was the survey design that collected data on L1 and L2 background of participants. Although the study did elicit data for these variables, the backgrounds of participants were far more complicated than anticipated. Consequently, the data obtained were not as accurate or informative as expected. This lack of detail in documenting the L2 learning experience is also evident in the study design, which conflated L2 learning in and out of the classroom with training in LAAs as well as familiarity with the LLAMA tests. In summary, however, despite these challenges and limitations, the findings of the current study offer some important insights and questions to the field, which will be highlighted in the final chapter.
CHAPTER 7: CONCLUSION

The current study was motivated by a desire to improve the L2 aptitude of L2 learners and in so doing to make a valuable contribution to L2 aptitude theory and research. To achieve these objectives, the current study focused on the analysis of the issues of stability and its correlate, untrainability, in L2 aptitude. The study extended previous research into these issues by pre- and post-testing across a shorter period of time (8 weeks as opposed to 12 months; e.g. Sáfár & Kormos, 2008; Ganschow & Sparks, 1995) and by investigating adult L2 learners (as opposed to high school aged learners). The current study was also the first in this area to quantify effect sizes of gain scores and include a comparative analysis of sub-groups with below and above average scores on L2 aptitude pre-tests. The study investigated the trainability of L2 aptitude, specifically language analytic abilities, for the first time in almost 50 years and extended the scope of the two previous training studies by implementing a regression analysis to ascertain the relative importance of training compared with L2 aptitude pre-test predictors. The study pre-tested L2 aptitude before L2 classes began and post-tested 8-9 weeks later. In the interim, participants completed a 6-week online training course in language analytic abilities, with no control group included. The analyses used in the current study resulted in a richer understanding of which individuals achieved the greatest gains in L2 aptitude, as well as which specific aspects of the training course had an effect on post-test scores of language analytic abilities.

The current chapter presents the major conclusions of the study and discusses their implications for the field of L2 aptitude research. The current study’s limitations are then
identified, with the chapter concluding by proposing directions for future research into the issue of stability and trainability in L2 aptitude.

7.1 Main findings of the study

In general, the study’s findings support the view of L2 aptitude as dynamic and sensitive to experience (see section 3.1.2) and challenge the assumptions of stability and untrainability (see section 3.1.1). Specifically, the study had three major findings, each of which is discussed below.

First, the current study found that in contrast to traditional assumptions, L2 aptitude is not stable. This finding was based on the results of two separate analyses. The results of paired-sample \( t \)-tests for pre- and post-tests of all LLAMA tests showed significant gains for all tests except the LLAMA E test, which showed a ceiling effect on both pre- and post-tests. This result clearly showed that the cohort in this study improved their L2 aptitude test scores over an 8 week period. Results from correlational analysis of all LLAMA pre- and post-tests supported the results from the paired-sample \( t \)-tests. Correlations between all pre- and post-tests were significant but weak, and below accepted levels of test-retest reliability. These results suggest that the test scores were not stable over the period, in line with the results from the paired-samples \( t \)-tests. Both sets of results challenge the traditional assumption of L2 aptitude as stable and, correlative, untrainable.

Second, in addition to finding that L2 aptitude is not stable, the current study found that gains in L2 aptitude varied across tests and for individuals with below and above average levels of L2 aptitude. Effect sizes for the paired-samples \( t \)-tests at the cohort level showed that gains were different across each of the LLAMA tests. Specifically, gains were strongest for
associative memory for vocabulary learning (LLAMA B) and for L2 aptitude as a whole (LLAMA Total), with gains of a medium effect size. Gains were weakest but still significant for sound discrimination for streams of speech sounds (LLAMA D) and for grammatical inferencing as part of language analysis (LLAMA F), with gains of a small effect size. This result, interpreted through Skehan’s (2002) model of SLA stages and L2 aptitude, might suggest that the L2 learning experience of memorising vocabulary and decoding the phonological system exerted a type of ‘training’ effect, through repetitive practice, on L2 aptitude. This view is also supported by the fact that language analytic abilities are less implicated in the initial stages of SLA. Understandably, gains were smallest for this area, because these abilities were less ‘practiced’. However, the results of the paired-sample $t$-tests conducted at the sub-group level of below and above average pre-test scores suggest a different interpretation. In general, participants with low pre-test scores made significant gains over the period, whereas participants with high-pretest scores did not. Moreover, the effect sizes for gain scores for the ‘low’ group were approximately the same across the different tests. This suggests that it is individuals with low levels of L2 aptitude who are most likely to significantly improve over time. This is to be expected, perhaps, as it is these individuals who have the most potential to increase their levels of L2 aptitude. The finding that significant gains were made mainly by individuals with low levels of L2 aptitude also suggests that the gains recorded are not merely an effect of familiarity with the test. If familiarity with the test was primarily responsible for the gains from pre- to post-test scores, we would expect to see at least comparable gains in the high level group, which we do not. The final point to note from these results concerns the gains for L2 aptitude as a whole. The results suggest that participants tended to generally have a weakness in only one area, but that

---

Note that no significant gains were found for the sound-symbol test of the LLAMA E for any of the paired-sample $t$-tests at any level.
overall they had high levels of L2 aptitude. Thus, the significant gains in at least one component of L2 aptitude appear to have been the cause for the significant increases in overall L2 aptitude, for both the ‘low’ and ‘high’ group. In short, while the current study found that individuals did improve their L2 aptitude test scores over time, closer analysis shows that these gains are more evident for individuals with low levels of L2 aptitude, and that this ‘weakness’ exists normally in merely one component of L2 aptitude.

Third, language analytic abilities increased significantly over the course of the study, with results indicating that the training intervention was at least partially responsible for these gains. Results from a random forest regression model and a stepwise linear model both included training variables as important predictors of post-test scores of language analytic abilities, showing that the training had some effect. Interestingly, in the random forest regression, the five training variables were more predictive of post-test scores than the pre-test. However, both models only explained a small amount of variance in the post-test scores of LAAs, which is most likely a result of the large variance in post-test scores and/or outside factors being responsible for the changes in post-test scores. In short, the current study found that LAAs, as a proxy for L2 aptitude, were trainable to an extent, and challenge the traditional assumption of the untrainability of L2 aptitude.

In summary, by means of its training course, the current study was able to investigate the stability of L2 aptitude and the trainability of LAAs, and, by so doing, to contribute to the understanding of L2 aptitude models. The current study extends the findings of previous studies by showing that gains in L2 aptitude do not seem attributable to maturational effects, as participants in the current study were adults. The current study was also the first to explicitly and exclusively investigate the efficacy of training in LAAs, and showed that
aspects of the training did indeed seem predictive of gains achieved. Again, these findings indicate that L2 aptitude is not stable, that LAAs are trainable to an extent, but that not everyone is expected to profit equally from L2 learning, instruction, or L2 aptitude training.

7.2 Implications of findings

In general, the findings from the current study suggest that L2 aptitude tests measure more than stable first-order abilities. This finding has important implications for the field in terms of both theory and practice.

In terms of theory, if L2 aptitude tests do measure more than stable first-order abilities, as the results indicate, then explanations of the construct need to expand beyond a focus on first-order abilities. Cognitive models of L2 aptitude, in particular, offer more holistic explanations of L2 aptitude. These models construe L2 aptitude as a composite of cognitive-based abilities in a hierarchical system. Importantly, these models view the learner as adaptive and profiting from feedback, gaining expertise through experience by acquiring metacognitive strategies and knowledge of L2 acquisition processes. These gains in higher-order abilities allow the learner to increase her or his effectiveness in L2 learning. The current study suggests that L2 aptitude tests implicitly measure these higher-order abilities of L2 aptitude, and it is these aspects of L2 aptitude that improve over time with experience. In short, the cognitive models’ prediction that L2 aptitude will be sensitive to experience is reflected in the results from this study. In addition, the cognitive models also predict that these higher-order abilities are teachable. This prediction was similarly borne out by results of this study.
The view of L2 aptitude as sensitive to experience and trainable has important implications for L2 aptitude research, and more specifically, L2 aptitude measurement. First, the question of when to test L2 aptitude becomes important, as does documenting the L2 learning experience of participants. In practice, the time of testing would of course be dependent on the focus of the investigation, e.g., testing monolinguals' L2 aptitude with static one-off testing should take place before L2 learning, as Carroll (1962, 1981) seems to suggest. However, if L2 aptitude is a dynamic variable, then static one-off testing utilising current L2 aptitude tests is unlikely to be the most appropriate choice. Instead, dynamic tests of L2 aptitude or reiterative testing regimes may offer a better alternative. For example, in experimental laboratory studies of SLA, pre- and post-testing L2 aptitude components might offer insight into the ability of participants to profit from instruction, feedback, or other variables of interest in SLA. Inevitably, new methods of L2 aptitude will also need to be accompanied by more sophisticated methods of statistical analysis.

Reconceptualising L2 aptitude as consisting of more than simply stable first-order abilities will also mean that SLA research and L2 teaching will need to determine where their focus should lie: on first-order abilities, metacognitive strategies, knowledge of L2 learning processes, or a combination of all three. For example, L2 teaching might explore ways of instructing L2 learners not only in what to learn, but also in how to learn the target language. Importantly, the learner’s L2 learning experience would also have to be taken into account. For example, factors that warrant consideration could include whether the individual is a monolingual, whether they have attained high levels of proficiency, and how much metalinguistic knowledge they possess. Inevitably, researchers and teachers alike will need to select those aspects of L2 experience that are most relevant to their objectives.
In short, the implications of L2 aptitude tests measuring more than stable first-order abilities are many and far ranging. It will be extremely interesting to see how L2 aptitude theory and research develop into the future.

7.3 Limitations of the research

The findings from the current study are promising, but its limitations must also be recognised. The study has many of the normal limitations to its scope, such as the sample, the context in which the study took place, and the backgrounds of its participants. However, the most important limitations to the current study are worth discussing in more detail. Significantly, the current study was an intervention in a real L2 learning situation. Although this results in higher ecological validity, this comes at the expense of internal validity. Consequently, the measurement of L2 aptitude (i.e., the LLAMA tests) and the LAAs training course were not optimal. The resulting data were messy, which led to complications in the analysis, such as the requirement to deal with a relatively large number of missing values. Additionally, the sample was one of convenience and not truly random. Lastly, it is also worth noting that the study conflated L2 learning in and out of the classroom with training in LAAs as well as familiarity with the LLAMA tests, because of methodological complications that resulted in a lack of detail in documenting the L2 learning experience. Overall, these factors comprised the strongest limitations to the current study.

7.4 Recommendations for future work

The findings from this study suggest new opportunities for research. Some of these potential opportunities are identified in this section.
As the current study was the first L2 aptitude training study in over 40 years, replication of its findings would be a significant contribution. Specific areas for a replication study to address could include: investigating over a longer timeframe; measuring L2 aptitude more than twice; sampling L2 learners with different L1 backgrounds and/or learning different L2s; and incorporating multiple control groups into the study design. Different control designs could include (1) an experimental group that undertakes pre- and post-testing of L2 aptitude, L2 learning, and L2 aptitude training; (2) a first control group that undertakes pre- and post-testing of L2 aptitude and L2 learning only; (3) and a second control group that undertakes pre- and post-testing L2 aptitude with no L2 learning or L2 aptitude training. This would allow for the different effects of the various types of experience to be teased out. Other studies could also investigate the training of components of L2 aptitude other than LAAs.

Perhaps the most important area for future research to focus on is L2 aptitude testing. Clearly, dynamic testing such as that implemented in the CANAL-FT (Grigorenko et al., 2000) warrant more research, as do iterative testing methodologies such as those researched by Krug et al. (2002). Fundamental research into behaviours of L2 aptitude test-taking offer the greatest potential for new insights into the construct. For example, verbal protocols and eye-tracking studies have been used in other areas of aptitude (e.g. Kyllonen, Lohman, & Snow, 1984) in cognitive psychology studies and might offer a model to follow for L2 aptitude research. If traditional L2 aptitude tests do measure more than stable first-order abilities, research is urgently needed to identify precisely what other aspects of L2 aptitude it is that these tests measure.

Finally, the current study also indicates the need for updated explanations of L2 aptitude. Any such explanations should reconstrue the construct as dynamic, sensitive to experience, and
amenable to training. As part of this, new tests of L2 aptitude and its components are most likely needed. These tests should include measures that isolate stable first-order abilities as well as quantify the effects of metacognitive strategies and knowledge of L2 learning processes.

Finally, the current study suggests some orientations for future practice in L2 learning and teaching. The results of the current study indicate that the skills of planning, monitoring, and reviewing can be beneficial to L2 learning. Consequently, L2 teaching might also include goals to improve L2 learning skills by explicitly instructing L2 learners in not only what to learn, but also how to learn the target language.
APPENDIX A: PROCEDURE FOR THE SYSTEMATIC REVIEW OF L2 APTITUDE RESEARCH

A.1 Objectives

This systematic review has two main objectives:

1. Gain an overall picture of L2 aptitude research as a field, identifying possible biases and problems in the area.
2. Identify areas in the L2 aptitude research that remain unclear or under-researched.

A.2 Process

The current systematic review was conducted following the guidelines set out in Pickering and Byrne (2014). To be included in the review, each paper had to meet all four of the following criteria: (1) be a journal article, PhD thesis, or a report; (2) pertain to empirical research that included some aspect of L2 aptitude; (3) utilise an instrument explicitly designed to measure L2 aptitude or one of its components; and (4) be published in English. Book chapters were excluded as a resource, because ensuring a comprehensive and exhaustive search of all the empirical studies available was not practically possible.

Scholarly electronic databases were searched to find empirical studies. These databases included: ERIC, JSTOR, ProQuest, SAGE, Web of Science, Wiley. These databases were searched between 07 May 2016 and 28 November 2016. For each database, an iterative keyword search was conducted. Search terms (see table A.1) were a combination of a keyphrase from (1) with the keyphrase from (2). The keywords and phrases from (1) identified: (a) L2 aptitude in general, (b) individual factors and abilities, and (c) measures and...
tests. The keywords and phrases from (2) delimited the field of research to second language learning. The keyphrases in table A.1 are written exactly as they were used in the keyword search, including punctuation and capitalisation. For example, the first search for each database was `aptitude AND ("second language learning" OR "second language acquisition" OR "foreign language learning")`, which was constructed by combining (1)(a) with (2).

**Table A.1. Keywords used in database search**

<table>
<thead>
<tr>
<th>Keyphrase</th>
<th>1.</th>
<th>2.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>aptitude</td>
<td>AND (&quot;second language learning&quot; OR &quot;second language acquisition&quot; OR &quot;foreign language learning&quot;)</td>
</tr>
<tr>
<td>b.</td>
<td>(&quot;phonetic coding&quot; OR &quot;phonemic coding&quot;)</td>
<td>(&quot;sound-symbol association&quot; OR &quot;sound discrimination&quot;)</td>
</tr>
<tr>
<td></td>
<td>(&quot;language analytic ability&quot;)</td>
<td>(&quot;grammatical sensitivity&quot; OR &quot;inductive language learning&quot;)</td>
</tr>
<tr>
<td></td>
<td>(&quot;sound-symbol association&quot; OR &quot;sound discrimination&quot;)</td>
<td>(&quot;associative memory&quot; OR &quot;rote memory&quot;)</td>
</tr>
<tr>
<td></td>
<td>(&quot;working memory&quot; AND aptitude)</td>
<td></td>
</tr>
<tr>
<td>c.</td>
<td>(&quot;modern language aptitude test&quot; OR &quot;MLAT&quot;)</td>
<td>(&quot;pimsleur language aptitude battery&quot; OR &quot;PLAB&quot;)</td>
</tr>
<tr>
<td></td>
<td>(&quot;defense language aptitude battery&quot; OR &quot;DLAB&quot;)</td>
<td>(&quot;army language aptitude battery&quot; OR &quot;ALAT&quot;)</td>
</tr>
<tr>
<td></td>
<td>(&quot;VORD&quot; OR &quot;york language aptitude test&quot;)</td>
<td>(&quot;cognitive ability for novelty in acquisition of language&quot; OR &quot;CANAL-F&quot;)</td>
</tr>
<tr>
<td></td>
<td>(&quot;LLAMA&quot; AND test)</td>
<td>(&quot;LLAMA&quot; AND test)</td>
</tr>
<tr>
<td></td>
<td>(&quot;lunic language marathon&quot;)</td>
<td>(&quot;lunic language marathon&quot;)</td>
</tr>
</tbody>
</table>
A.2.1 Data items

Following the guidelines in Pickering and Byrne (2014), the following data categories were created: publication details, geographical information, research methods used, participant data, variables measured, the discipline of each study, and findings and conclusions.

For the first eight data items (publication details, geographical information, research methods used, participant data, variables measured, the discipline of each study), data was extracted from each paper according to the categories and definitions detailed in table A.3 (see section 2.4). Data for findings and conclusions were categorised by keywords into themes (see table A.2) extracted from five theoretical reviews: Dörnyei (2005), Robinson (2012), Sawyer & Ranta (2001), and Skehan (1998, 2002).

Table A.2. Themes and questions for L2 aptitude research

<table>
<thead>
<tr>
<th>Themes</th>
<th>Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose and relevance of L2 aptitude</td>
<td>Dörnyei (2005), Sawyer &amp; Ranta (2001), Skehan (1998)</td>
</tr>
<tr>
<td>The origins of L2 aptitude and its relationship with the L1</td>
<td>Dörnyei (2005), Sawyer &amp; Ranta (2001), Skehan (2002)</td>
</tr>
<tr>
<td>L2 aptitude and aptitude-treatment interactions (ATIs)</td>
<td>Dörnyei (2005), Robinson (2012), Sawyer &amp; Ranta (2001)</td>
</tr>
<tr>
<td>L2 aptitude, working memory, and other cognitive processes</td>
<td>Dörnyei (2005), Robinson (2012), Sawyer &amp; Ranta (2001)</td>
</tr>
<tr>
<td>L2 aptitude and intelligence</td>
<td>Dörnyei (2005), Sawyer &amp; Ranta (2001), Skehan (1998)</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

Figure A.1 below shows how many papers on L2 aptitude were published each year for the period of the review, 1962-2016.

![Figure A.1 Papers on L2 aptitude research published 1962-2016](image)

*Figure A.1 Papers on L2 aptitude research published 1962-2016*
A.2.2 Study selection

From the initial screening of the keyword database searches, 609 papers were selected and assessed for eligibility. From these 609 studies, 375 studies were excluded for being non-empirical (e.g. theoretical reviews, book reviews), book chapters, or not relating to the prediction of L2 learning. From the remaining 228 papers on the prediction of L2 learning, another 105 papers were excluded for not utilising an instrument explicitly designed to measure L2 aptitude or one of its components. The remaining 123 papers were included and constituted the dataset for the systematic review.

Figure A.2 Flowchart of screening process for systematic review.
A.2.3 Terminology

For the purpose of the review, the term “paper” refers to a publication and the term “study” refers to research where data was collected from a group of participants and used for analysis. A paper can report on more than one study, with each study being recorded separately in the review dataset, as long as the participants and the data differ between the studies. This situation arises where replication or follow-up studies are conducted. While follow-up studies may be considered part of a longitudinal data collection and therefore constitute a single study, this categorisation was not used, as the follow-up studies recorded almost always had a different research focus and/or different data, e.g. the use of a different instrument, and did not maintain the complete original sample of participants.

A.2.4 Data items and their definitions

For the data collection, a range of terms were used to capture all the data. Table A.3 is a complete list of all the data items tracked through the systematic review and how they were defined as part of the study.

Table A.3. Data items and their definitions

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliographic Data</td>
<td>Title</td>
<td>Name of the paper.</td>
</tr>
<tr>
<td></td>
<td>Author 1-3</td>
<td>First three listed authors of the paper in order.</td>
</tr>
<tr>
<td></td>
<td>Authors Additional</td>
<td>The remaining authors.</td>
</tr>
<tr>
<td></td>
<td>Year published</td>
<td>The year the paper was published, as written on the paper or else as per metadata on Zotero.</td>
</tr>
<tr>
<td></td>
<td>Name of publication</td>
<td>The name of the publication as written on the paper or else as per metadata on Zotero.</td>
</tr>
</tbody>
</table>
### Type of publication
Each publication was categorised as a peer reviewed journal article, a published PhD dissertation, or an official published report, e.g. for a government organisation. Official reports were identified as such from information on the covers or from metadata.

### Location

<table>
<thead>
<tr>
<th>Location</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher 1-3 institution</td>
<td>The institution for each author as per the details given on the paper.</td>
</tr>
<tr>
<td>City/Town 1-3</td>
<td>The city or town of the institution for each of the first three authors, taken from the paper if available or Google Maps if not.</td>
</tr>
<tr>
<td>State/Region 1-3</td>
<td>The state or region of the institution for each of the first three authors, taken from the paper if available or Google Maps if not.</td>
</tr>
<tr>
<td>Country 1-3</td>
<td>The country of the institution for each of the first three authors, taken from the paper if available or Google Maps if not.</td>
</tr>
</tbody>
</table>

### Participants

<table>
<thead>
<tr>
<th>Participants</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number</td>
<td>The total number of participants for who data was collected and analysed, as reported in the paper.</td>
</tr>
<tr>
<td>Average age</td>
<td>The average age for the total number of participants as reported or calculated, if appropriate.</td>
</tr>
<tr>
<td>Number of males/females</td>
<td>The total number of female and male participants, as reported in the paper.</td>
</tr>
<tr>
<td>SES</td>
<td>The socio-economic status of participants, if reported, categorised as lower, middle, upper, or not reported. No attempt was made to standardize the measures across studies, but rather categories were reported as per the study.</td>
</tr>
<tr>
<td>Education level</td>
<td>The highest level of education undertaken by participants at the time of the study, categorised as none, elementary, high school, university, post-graduate, or not reported. No attempt was made to standardize the measures across studies, but rather categories were reported as per the study.</td>
</tr>
<tr>
<td>L1s</td>
<td>The first language of participants as reported in the study. In the case where this was not reported, the context of the study, e.g. country, was used to...</td>
</tr>
</tbody>
</table>
infer the L1 and the entry was also explicitly recorded as inferred and not reported.

<table>
<thead>
<tr>
<th>Context</th>
<th>L2 learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>How participants were learning/had learned the L2 in regards to the study. Categories were Instructed where formal learning had taken place; Uninstructed where no formal learning had taken place; both where the study had compared language learning under these conditions separately, i.e. two disparate groups; and mixed where there was no distinction between groups or within participants for these two conditions; and not reported where no information was given.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of instruction</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>If participants had undertaken instruction, then the type of instruction was also categorised. Communicative = focus of instruction was skills related to communicating in the L2, especially speaking; Translation = where the L2 was learned through translation tasks or activities; Audio-lingual = as reported as audio-lingual instruction; Experimental = non-standard instruction, normally designed by (a member of) the research team; and not reported.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Setting</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The environment in which the second language was learned was categorised: classroom = a formal school environment, most often in a teacher-led class; informal setting = where learning took place outside of the classroom, e.g. watching movies, etc.; immersion = when an immersion program had been explicitly reported, whether it was genuine or artificial, e.g. CLIL; laboratory = when participants underwent an experimental learning procedure run in a (computer) laboratory; mixed = when the participant group included individuals from multiple settings and no distinction between them was made.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data</th>
<th>Type of data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The type of data collected was categorised according to Dörnyei (2007): Quantitative = numerical data representing predetermined logical scales of values, e.g. aptitude test; Qualitative = verbal categories of data determined by examining the data after collection; Mixed = where both these types of</td>
</tr>
<tr>
<td>Collection</td>
<td>The collection of data was also categorised according to Dörnyei (2007): Longitudinal = data collected at two different points in time from the same participants to explain a change over time; Cross-sectional = data collected at a single point in time that is not longitudinal; Both = where both types of collection took place, e.g. a background survey followed pre- and post-testing.</td>
</tr>
<tr>
<td>Location 1-4 of data collection</td>
<td>The country in which the data was collected as reported in the study. Up to four locations were recorded for one study.</td>
</tr>
<tr>
<td>L2s</td>
<td>The second language being studied by the participants, whether it was the target language of the study or the language(s) being learned by participants. Artificial referred to studies where an modified language system was used, whether it be complete or not. Not reported referred to those studies that did not detail the second language being learned by participants. None referred to when no second language was being learned by participants.</td>
</tr>
<tr>
<td>L2 aptitude focus</td>
<td>L2 aptitude in study</td>
</tr>
<tr>
<td>L2 aptitude study</td>
<td>L2 aptitude focus</td>
</tr>
</tbody>
</table>
where tests of language analytic abilities, e.g. MLAT IV or LLAMA F, were used in the analysis; L1 skills = where measures of L1 abilities, e.g. word decoding, were explicitly labelled as coming under the same factor as L2 aptitude.

**Design**

The level of interaction with participants the researchers designed into the study, categorised as: experimental = where the study explicitly introduced a specific treatment and then compared these outcomes against a control; observational = where the study collected data from participants normal course of L2 learning.

**Aspect investigated**

Taken from theoretical reviews of L2 aptitude (see Carroll 1981, 1990; Dörnyei 2005; Dörnyei & Skehan 1998; Robinson 2002, 2012; Sawyer & Ranta 2001; Skehan 1990, 1998, 2002), each study was categorised as focussing on a certain aspect of L2 aptitude:
- test construct validity = where L2 aptitude measures were analysed with other related constructs, e.g. intelligence, to see if L2 aptitude truly exists independently;
- test predictive validity = where L2 aptitude measures were analysed for their ability to predict L2 learning outcomes;
- trainability = where participants undertook a treatment of some kind that aimed to enhance L2 aptitude abilities in some way;
- stability = where L2 aptitude was analysed over time to see if it changed or not;
- origin = where other variables were analysed to see if they were responsible for the development of L2 aptitude abilities;
- neuro-biological basis = where brain imaging data was analysed to investigate the loci of L2 aptitude abilities;
- learning treatment = where the validity of L2 aptitude was analysed according to the method of language learning participants underwent;
- structure of L2 aptitude = where L2 aptitude measures and possibly other measures were analysed, normally with principal components analysis or structural equation modelling, to investigate the nature of L2 aptitude; ATIs = where the method of language learning was analysed in relation to L2 aptitude and to see if different L2 aptitude abilities were related to different language learning approaches;
L2 aptitude + motivation + attitude model = where models, e.g. Gardner’s sociocultural model, were explicitly tested; test-retest reliability = where L2 aptitude measures were tested across multiple points in time and then analysed for their stability/reliability.

Dependent variables

The dependent variable was determined by the statistical test used for the analysis, e.g. the response variable in a general linear model. Categories covered: overall L2 achievement = where a score was given for L2 learning as a whole, e.g. end of course grade; specific L2 achievement = where a score was given for a selected area of L2 learning over a period of time, e.g. course grade for speaking; specific L2 skill = where a score is given for a test of a specific L2 learning ability, e.g. auditory perception test; specific L2 knowledge = where a score was given for a test for knowledge of something specific in the L2, e.g. recall of learned vocabulary or a grammatical rule of the L2; L2 aptitude scores = where a score was given for a measure of L2 aptitude, be it aptitude as a whole or a specific ability; age of arrival = where participants were analysed by groups determined from what age they arrived in the L2-speaking environment; dropout rate = where participants were analysed by groups determined by if they discontinued L2 learning or not; L2 anxiety = where participants were analysed by groups determined by a measure of anxiety for learning a second language, L2 experience and development = where the Foreign Language Screening Instrument for Colleges (Sparks & Ganschow 1991) or an adaptation was used as a measure of the participants’ language learning history; ordinal = where participants either belonged to an ordinal category or not for the purpose of analysis of variance; L1 skills = where scores from a variety of possible measures in L1 skills were used as a measure of first language knowledge and ability; teacher ratings = a prediction by the instructor/teacher of a course, be it on a scale or predicted grade of L2 achievement or learning facility; WM test scores = a score on any test of working memory, be it on processing, storage, or both.
<table>
<thead>
<tr>
<th>Named dependent variable</th>
<th>The name(s) of the specific variable from the general category of the dependent variable and the name of the measure(s)/instrument(s) used.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control variables</td>
<td>Any variables used to limit the sample of participants in the study, control for differences among different groups of participants, or used as a covariant.</td>
</tr>
<tr>
<td>Motivation as IV</td>
<td>A “yes” or “no” response to indicate if motivation was an explicit independent variable used in the analysis.</td>
</tr>
<tr>
<td>Measure of motivation</td>
<td>The name of the instrument(s) used to measure motivation.</td>
</tr>
<tr>
<td>Attitude as IV</td>
<td>A “yes” or “no” response to indicate if attitude was an explicit independent variable used in the analysis.</td>
</tr>
<tr>
<td>Measure of attitude</td>
<td>The name of the instrument(s) used to measure attitude.</td>
</tr>
<tr>
<td>Anxiety as IV</td>
<td>A “yes” or “no” response to indicate if anxiety was an explicit independent variable used in the analysis.</td>
</tr>
<tr>
<td>Measure of anxiety</td>
<td>The name of the instrument(s) used to measure anxiety.</td>
</tr>
<tr>
<td>WM as IV</td>
<td>One of five categories to indicate whether working memory was an explicit independent variable used in the analysis and if so, what aspect was measured. Categories consisted of: Processing = a measure of the ability to process in working memory, e.g. multisyllabic word test; Storage = a measure of the storage capacity of working memory, e.g. L1 PSTM nonword repetition; Both = a measure of the simultaneous storage and processing of working memory, e.g. L1 reading span test; No = where working memory was not an independent variable in the study.</td>
</tr>
<tr>
<td>Measure of WM</td>
<td>The name of the instrument(s) used to measure working memory.</td>
</tr>
<tr>
<td>General Cognition as IV</td>
<td>A “yes” or “no” response to indicate if general cognition was an explicit independent variable used in the analysis.</td>
</tr>
<tr>
<td>L2 aptitude itself (only)</td>
<td>Aspect of general cognition</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Measure of general cognition</td>
<td>The name of the instrument(s) used to measure general cognition.</td>
</tr>
<tr>
<td>Age as IV</td>
<td>A “yes” or “no” response to indicate if age was an explicit independent variable used in the analysis.</td>
</tr>
<tr>
<td>Age of arrival</td>
<td>A “yes” or “no” response to indicate if age of arrival was an explicit independent variable used in the analysis.</td>
</tr>
<tr>
<td>Personality as IV</td>
<td>A “yes” or “no” response to indicate if personality was an explicit independent variable used in the analysis.</td>
</tr>
<tr>
<td>Measure of personality</td>
<td>The name of the instrument(s) used to measure personality.</td>
</tr>
<tr>
<td>Gender as IV</td>
<td>A “yes” or “no” response to indicate if gender was an explicit independent variable used in the analysis.</td>
</tr>
<tr>
<td>Tolerance of ambiguity as IV</td>
<td>A “yes” or “no” response to indicate if tolerance of ambiguity was an explicit independent variable used in the analysis.</td>
</tr>
<tr>
<td>Measure of tolerance of ambiguity</td>
<td>The name of the instrument(s) used to measure tolerance of ambiguity.</td>
</tr>
<tr>
<td>Language learning styles as IV</td>
<td>A “yes” or “no” response to indicate if language learning styles was an explicit independent variable used in the analysis.</td>
</tr>
<tr>
<td>Measure of learning styles</td>
<td>The name of the instrument(s) used to measure learning styles.</td>
</tr>
<tr>
<td>Other IV</td>
<td>A named variable not contained in the list above, including academic achievement, L1 skills, metalinguistic knowledge, and zygosity (see #table for complete list).</td>
</tr>
<tr>
<td>Measure of other IV</td>
<td>The name of the instrument(s) used to measure the independent variables in “Other IV”.</td>
</tr>
<tr>
<td>Measure of L2 aptitude</td>
<td>The name of the specific instrument(s) used to measure L2 aptitude or L2 aptitude abilities. This includes the standardised and common tests of L2 aptitude, their adaptations (e.g. translations), or instruments explicitly designed to measure L2 aptitude or an L2 aptitude ability. (See #table for complete list.)</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Analysis approach</td>
<td>The design for the analysis of data used in the study, including the categories: between group comparison = where analysis of variance was used; within group comparison = where analysis of variance was used; variance explained = where regression was used; structural analysis = where principal components analysis or structural equation modelling were used; relationships between variables = where correlation was used.</td>
</tr>
<tr>
<td>Number of groups</td>
<td>The number of groups used in the analysis.</td>
</tr>
<tr>
<td>Division of groups</td>
<td>The principle behind the division of groups in the study. The major categories were: ability = where participants were grouped based on their ability on a certain task or procedure, e.g. L2 proficiency, general cognition, etc.; treatment = where participants were grouped based on what kind of experience they had, e.g. experimental instructional method vs control group; cohort = where participants were selected based on being part of a group, e.g. year 9 students in a particular school district; etc. (see #table for complete list).</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>The specific statistical tests used in the analysis of L2 aptitude. Includes both initial and follow-up testing, e.g. ANOVA followed up by Scheffé’s test. (see #table for complete list.)</td>
</tr>
<tr>
<td>Significance of results</td>
<td>The p-value for the main statistical analyses of L2 aptitude in the study. Categories were: non-significant = p-value &gt; .10; approaching significance = p.value between .05 and .10; significant = p-value between .05 and .01; highly significant = p.value &lt; .01; not reported = no p-value mentioned or no statistical tests carried out.</td>
</tr>
<tr>
<td>Effect size</td>
<td>The effect size for the statistical test, for Cohen’s d (d), correlation coefficients (r), and eta-squared (( \eta^2 )). Categories were taken following</td>
</tr>
</tbody>
</table>
Larson-Hall (2013): small = \( d: .40, r: .25, \eta^2: .01; \) medium = \( d: .70, r: .40, \eta^2: .06; \) large = \( d: 1.0, r: .60, \eta^2: .14; \) very large = \( d: >1.0, r: >.60, \eta^2: >.14; \) not reported = no effect size mentioned or no statistical tests carried out.

<table>
<thead>
<tr>
<th>Findings</th>
<th>Reporting of major findings of analysis as relate to L2 aptitude.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conclusions</td>
<td>Reporting of major conclusions of study as relate to L2 aptitude.</td>
</tr>
<tr>
<td>L2 aptitude as ID (only)</td>
<td>Use of L2 aptitude How L2 aptitude was included into the study, as categorised by: group assignment = where participants were grouped according to a score on L2 aptitude, e.g. in ATI-factorial ANOVA studies; explaining results = where L2 aptitude scores were used to help explain results between or within individuals; interaction with other IDs = where L2 aptitude was investigated as interacting with another variable of interest, e.g. L2 aptitude and L2 anxiety; [Other…]]</td>
</tr>
<tr>
<td>L2 aptitude as control (only)</td>
<td>Investigation of study A short description of what the study was investigating, particularly in relation to the use of L2 aptitude</td>
</tr>
<tr>
<td>L2 aptitude as control (only)</td>
<td>Use of L2 aptitude How L2 aptitude was used within the study, as categorised by: group assignment = where participants were placed into groups based on an L2 aptitude score; equivalence between groups = where L2 aptitude was used as a covariant in the analysis; explaining results = as above; interaction with other IDs = as above.</td>
</tr>
</tbody>
</table>
APPENDIX B: REFERENCE LIST FOR PAPERS IN SYSTEMATIC REVIEW OF L2 APTITUDE


Bain, S. K., McCallum, R. S., Bell, S. M., Cochran, J. L., & Sawyer, S. C. (2010). Foreign Language Learning Aptitudes, Attitudes, Attributions, and Achievement of


APPENDIX C: SUPPLEMENTARY RESULTS

The following appendix contains all the essential elements of the full analysis. In particular, all results from the stepwise linear models are reported.

The following is the complete list of libraries needed to run all analyses.

```r
knitr::opts_chunk$set(message = FALSE, warning = FALSE)

library(MASS)
library(knitr)
library(readxl)
library(psych)
library(ggplot2)
library(gridExtra)
library(moments)
library(robustbase)
library(lme4)
library(randomForest)
library(dplyr)
library(tidyr)
```
C.1 Introduction

This chapter details all results from all statistical analyses carried out in investigating the research questions. All statistical analyses were carried out in R 3.3.2 (R Core Team, 2016), using an alpha level of 0.05 for all relevant tests.

The results are organized by research question.

The first research question contains the following steps in the analysis:

1. assumptions of the paired-samples t-test
2. differences between participants based on background variables
3. descriptive statistics for the LLAMA tests
4. paired-samples t-tests at the cohort level
5. paired-samples t-tests for sub-groups of below vs above average pre-test scores
6. correlations between all pre- and post-test scores
7. The second research question contains the following steps in the analysis:
8. data preparation of all independent variables
   i. trimmed times of all training tasks and LLAMA tests
   ii. L2 aptitude pre-test scores
   iii. training course completion
   iv. training course approach
   v. training task accuracy
   vi. training task speed
9. random forest model
10. stepwise linear regression model
   i. correlations of predictors with response
C.2 Research Question 1: Are post-test scores for L2 aptitude, operationalised as the LLAMA tests, significantly different to pre-test scores?

C.2.1 Assumptions of paired-samples t-tests

Parametric paired-samples t-tests were carried out to check if the difference between pre- and post-test scores from the same participants was statistically significant. T-tests were two-tailed to factor in the likelihood that some participants would experience negative gains as a result of an inhibitory effect of instruction, either from the L2 Spanish course or the language analytic abilities training course (see chapter 3.1.3.4 for discussion). The assumptions of paired-samples t-tests are that gain scores between tests are normally distributed and there are no outliers. Visual checks for normality for all LLAMA tests include histograms, QQ (quantile-quantile) plots, and boxplots.

Table C.1 Skewness, kurtosis, and Shapiro-Wilks for LLAMA tests

<table>
<thead>
<tr>
<th>LLAMA tests</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>W (Shapiro-Wilks)</th>
<th>p (Shapiro-Wilks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.41</td>
<td>0.86</td>
<td>0.98</td>
<td>0.18</td>
</tr>
<tr>
<td>D</td>
<td>0.23</td>
<td>0.73</td>
<td>0.97</td>
<td>0.02</td>
</tr>
<tr>
<td>E</td>
<td>-1.06</td>
<td>3.22</td>
<td>0.88</td>
<td>0.00</td>
</tr>
<tr>
<td>F</td>
<td>-0.41</td>
<td>-0.18</td>
<td>0.97</td>
<td>0.07</td>
</tr>
<tr>
<td>Total</td>
<td>1.06</td>
<td>2.55</td>
<td>0.93</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Figure C.1 Histograms of LLAMA tests B, D, E, and F

Figure C.2 Histogram of LLAMA Total
Figure C.3 QQ plots for LLAMA tests B, D, E, F

Figure C.4 QQ plot for LLAMA Total
Figure C.5 Boxplots for LLAMA B, D, E, F

Figure C.6 Boxplot for LLAMA Total
**C.2.2 Background variables of participants**

Before running any inferential statistical tests, an independent factorial ANOVA was run for each LLAMA test and LLAMA Total for age, L1, and L2 experience level (see chapter 4.2 for summary descriptions of each variable) to determine if significant differences existed between for these individual factors in the sample. L1 was a binary factor of English vs. non-English, which was relevant to the study, as English was the language of instruction and many participants were not L1 speakers of English. The L2 experience level factor attempted to capture experience in the process of learning second languages, as opposed to merejust proficiency or number of languages learned, and consisted of four levels (see chapter 4.2 for descriptions of these levels).

```r
## Response LLAMA_B_pre :
## Df Sum Sq Mean Sq F value  Pr(>F)
## Age     1  1869 1869.13  4.0814 0.04581 *
## L1      1    14  13.88  0.0303 0.86212
## L2_Level 3  537  179.17  0.3912 0.75955
## Residuals 109 49918  457.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## Response LLAMA_D_pre :
## Df Sum Sq Mean Sq F value  Pr(>F)
## Age     1   131 131.30  0.4423 0.5074
## L1      1   327 327.49  1.1033 0.2959
## L2_Level 3  650  216.70  0.7300 0.5362
```
## Residuals 109 32355 296.84

## Response LLAMA_E_pre:

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1</td>
<td>611</td>
<td>611.01</td>
<td>1.3138</td>
</tr>
<tr>
<td>L1</td>
<td>1</td>
<td>1352</td>
<td>1352.00</td>
<td>2.9071</td>
</tr>
<tr>
<td>L2_Level</td>
<td>3</td>
<td>1707</td>
<td>568.87</td>
<td>1.2232</td>
</tr>
<tr>
<td>Residuals</td>
<td>109</td>
<td>50693</td>
<td>465.07</td>
<td></td>
</tr>
</tbody>
</table>

---

## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Response LLAMA_F_pre:

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1</td>
<td>741</td>
<td>741.36</td>
<td>1.1542</td>
</tr>
<tr>
<td>L1</td>
<td>1</td>
<td>621</td>
<td>621.30</td>
<td>0.9673</td>
</tr>
<tr>
<td>L2_Level</td>
<td>3</td>
<td>6346</td>
<td>2115.18</td>
<td>3.2931</td>
</tr>
<tr>
<td>Residuals</td>
<td>109</td>
<td>70012</td>
<td>642.32</td>
<td></td>
</tr>
</tbody>
</table>

---

## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Response pre_total:

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1</td>
<td>7009</td>
<td>7009.2</td>
<td>2.3027</td>
</tr>
<tr>
<td>L1</td>
<td>1</td>
<td>687</td>
<td>687.2</td>
<td>0.2258</td>
</tr>
</tbody>
</table>
## L2_Level      3  15951  5317.1  1.7468 0.1617
## Residuals   109 331788  3043.9

## Single term deletions

## Model:
## LLAMA_B_pre ~ Age + L1 + L2_Level
##
## Df Sum of Sq RSS AIC F Value Pr(F)
## <none>                49918 710.42
## Age       1    1659.60 51577 712.18  3.6239 0.05959 .
## L1        1     25.15 49943 708.48  0.0549 0.81515
## L2_Level  3    537.50 50455 705.65  0.3912 0.75955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Single term deletions

## Model:
## LLAMA_B_pre ~ Age
##
## Df Sum of Sq RSS AIC F Value Pr(F)
## <none>              50469 703.68
## Age     1    1869.1 52338 705.86   4.185 0.04311 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Single term deletions

### Model:

### LLAMA_F_pre ~ Age + L1 + L2_Level

### Df Sum of Sq RSS AIC F Value Pr(F)
### <none> 70012 749.32
### Age 1 770.8 70783 748.58 1.2000 0.27573
### L1 1 2295.4 72308 751.03 3.5736 0.06136 .
### L2_Level 3 6345.5 76358 753.30 3.2931 0.02333 *
### ---

### Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Single term deletions

### Model:

### LLAMA_F_pre ~ L1 + L2_Level

### Df Sum of Sq RSS AIC F Value Pr(F)
### <none> 70783 748.58
### L1 1 2112.4 72896 749.96 3.2828 0.07274 .
### L2_Level 3 6466.3 77249 752.63 3.3496 0.02170 *
### ---

### Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Tukey multiple comparisons of means

### 95% family-wise confidence level
Table C.2 LLAMA F pre-test scores for English vs non-English L1 participants for differing levels of L2 experience

<table>
<thead>
<tr>
<th>L1</th>
<th>None</th>
<th>Low</th>
<th>Some</th>
<th>Extensive</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>48.49</td>
<td>52.86</td>
<td>52.50</td>
<td>75.00</td>
</tr>
<tr>
<td>Other</td>
<td>NA</td>
<td>NA</td>
<td>38.57</td>
<td>62.00</td>
</tr>
</tbody>
</table>

### Fit: aov(formula = LLAMA_F_pre ~ L1 + L2_Level, data = check.parts)

### $L1$

<table>
<thead>
<tr>
<th></th>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other-English</td>
<td>-4.561444</td>
<td>-15.12595</td>
<td>6.003064</td>
<td>0.3940415</td>
</tr>
</tbody>
</table>

### $L2$ Level

<table>
<thead>
<tr>
<th></th>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-None</td>
<td>2.758197</td>
<td>-22.1254606</td>
<td>27.64185</td>
<td>0.9915328</td>
</tr>
<tr>
<td>Some-None</td>
<td>-1.748356</td>
<td>-16.1928102</td>
<td>12.69610</td>
<td>0.9890422</td>
</tr>
<tr>
<td>Extensive-None</td>
<td>20.480657</td>
<td>0.8695874</td>
<td>40.09173</td>
<td>0.0370859</td>
</tr>
<tr>
<td>Some-Low</td>
<td>-4.506552</td>
<td>-30.6648569</td>
<td>21.65175</td>
<td>0.9696078</td>
</tr>
<tr>
<td>Extensive-Low</td>
<td>17.722460</td>
<td>-11.6068542</td>
<td>47.05177</td>
<td>0.3963099</td>
</tr>
<tr>
<td>Extensive-Some</td>
<td>22.229012</td>
<td>1.0239533</td>
<td>43.43407</td>
<td>0.0360912</td>
</tr>
</tbody>
</table>
### Descriptive statistics

Descriptive statistics for all tests were calculated for all tests. The mean, range, standard deviation, variance, and 95% confidence interval for each LLAMA pre- and post-test are all calculated.
### Table C.3 Descriptive statistics for all LLAMA pre- and post-tests

<table>
<thead>
<tr>
<th>LLAMA test</th>
<th>n</th>
<th>Mean</th>
<th>Range</th>
<th>Standard deviation</th>
<th>Variance</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>B (pre)</td>
<td>90</td>
<td>50.83</td>
<td>0-100</td>
<td>21.47</td>
<td>460.81</td>
<td>46.34-55.33</td>
</tr>
<tr>
<td>B (post)</td>
<td>90</td>
<td>67.89</td>
<td>0-100</td>
<td>25.49</td>
<td>649.99</td>
<td>62.55-73.23</td>
</tr>
<tr>
<td>D (pre)</td>
<td>90</td>
<td>29.94</td>
<td>0-70</td>
<td>17.05</td>
<td>290.73</td>
<td>26.37-33.52</td>
</tr>
<tr>
<td>D (post)</td>
<td>90</td>
<td>37.56</td>
<td>0-75</td>
<td>17.36</td>
<td>301.26</td>
<td>33.92-41.19</td>
</tr>
<tr>
<td>E (pre)</td>
<td>85</td>
<td>82.12</td>
<td>10-100</td>
<td>19.65</td>
<td>385.94</td>
<td>77.88-86.36</td>
</tr>
<tr>
<td>E (post)</td>
<td>85</td>
<td>84.76</td>
<td>0-100</td>
<td>25.28</td>
<td>639.23</td>
<td>79.31-90.22</td>
</tr>
<tr>
<td>F (pre)</td>
<td>90</td>
<td>49.56</td>
<td>0-100</td>
<td>26.39</td>
<td>696.43</td>
<td>44.03-55.08</td>
</tr>
<tr>
<td>F (post)</td>
<td>90</td>
<td>56.70</td>
<td>0-100</td>
<td>27.85</td>
<td>775.38</td>
<td>50.87-62.53</td>
</tr>
<tr>
<td>Total (pre)</td>
<td>85</td>
<td>212.41</td>
<td>25-330</td>
<td>50.55</td>
<td>2555.72</td>
<td>201.51-223.32</td>
</tr>
<tr>
<td>Total (post)</td>
<td>85</td>
<td>248.33</td>
<td>30-345</td>
<td>64.30</td>
<td>4134.72</td>
<td>234.46-262.20</td>
</tr>
</tbody>
</table>

### C.2.4 Paired-samples t-tests

A two-tailed paired-samples t-test was carried out for each LLAMA test to find any significant differences between pre- and post-test scores. Cohen’s d was also calculated as a measure of effect size, that is, how big the difference was between pre- and post-test scores.
Table C.4 Results from paired-sample t-tests, 95% confidence intervals, and effect sizes (Cohen's d) for each LLAMA test (B, D, E, F, and Total)

<table>
<thead>
<tr>
<th>LLAMA test</th>
<th>df</th>
<th>t</th>
<th>p</th>
<th>95% CI</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>89</td>
<td>7.32</td>
<td>0.000</td>
<td>12.43-21.68</td>
<td>0.61</td>
</tr>
<tr>
<td>D</td>
<td>89</td>
<td>4.10</td>
<td>0.000</td>
<td>3.92-11.30</td>
<td>0.40</td>
</tr>
<tr>
<td>E</td>
<td>84</td>
<td>1.01</td>
<td>0.317</td>
<td>-2.59-7.88</td>
<td>0.11</td>
</tr>
<tr>
<td>F</td>
<td>89</td>
<td>2.07</td>
<td>0.041</td>
<td>0.30-13.99</td>
<td>0.21</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>5.57</td>
<td>0.000</td>
<td>23.10-48.73</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Figure C.7 Boxplots for pre- and post-tests for LLAMA B, D, E, F
Figure C.8 Boxplots for pre- and post-tests for LLAMA Total

Figure C.9 Scatter plots for LLAMA B, D, E, F with the regression line (red) and its 95% confidence interval (blue shading)
C.2.4.1 Paired-sample t-tests for low and high groups

Paired-samples t-tests were also carried out for all LLAMA tests with Low and High pre-test score groupings. A score was classified as Low if it was considered an average score or below as per the LLAMA test manual (Meara 2005), and High if was considered above this average. The purpose of analysing Low and High pre-test score groups was to check if only those with initially low L2 aptitude were making the greater gains, as they had the most possibility to improve (Ganschow & Sparks, 1995; Politzer & Weiss, 1969; Sparks et al., 1998; cf. Yeni-Komshian, 1965). Numerical results are reported in table 5.5 with t-test statistics, 95% confidence intervals, and effect sizes (Cohen’s d) for each group (Low and High) for each LLAMA test (B, D, E, F, and Total).

Figure C.10 Scatter plot for LLAMA Total with the regression line (red) and its 95% confidence interval (blue shading)
Table C.5 Results from paired-sample t-tests for Low and High groups for all LLAMA tests

<table>
<thead>
<tr>
<th>LLAMA test</th>
<th>df</th>
<th>t</th>
<th>p</th>
<th>95% CI</th>
<th>Effect size (Cohen’s d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low B</td>
<td>37</td>
<td>5.02</td>
<td>0.000</td>
<td>13.34-31.40</td>
<td>0.64</td>
</tr>
<tr>
<td>Low D</td>
<td>56</td>
<td>5.55</td>
<td>0.000</td>
<td>7.96-16.96</td>
<td>0.60</td>
</tr>
<tr>
<td>Low E</td>
<td>4</td>
<td>0.95</td>
<td>0.397</td>
<td>-28.90-58.90</td>
<td>0.43</td>
</tr>
<tr>
<td>Low F</td>
<td>38</td>
<td>5.51</td>
<td>0.000</td>
<td>15.63-33.76</td>
<td>0.67</td>
</tr>
<tr>
<td>Low Total</td>
<td>20</td>
<td>7.21</td>
<td>0.000</td>
<td>53.73-97.51</td>
<td>0.85</td>
</tr>
<tr>
<td>High B</td>
<td>51</td>
<td>5.81</td>
<td>0.000</td>
<td>0.00-17.73</td>
<td>0.63</td>
</tr>
<tr>
<td>High D</td>
<td>32</td>
<td>-0.28</td>
<td>0.783</td>
<td>0.78-4.80</td>
<td>0.05</td>
</tr>
<tr>
<td>High E</td>
<td>79</td>
<td>0.71</td>
<td>0.478</td>
<td>0.48-7.10</td>
<td>0.08</td>
</tr>
<tr>
<td>High F</td>
<td>50</td>
<td>-1.51</td>
<td>0.138</td>
<td>0.14-2.09</td>
<td>0.21</td>
</tr>
<tr>
<td>High Total</td>
<td>63</td>
<td>3.20</td>
<td>0.002</td>
<td>0.00-37.19</td>
<td>0.37</td>
</tr>
</tbody>
</table>
Figure C.11 Scatter plots for Low (triangle) and High (circle) groups for LLAMA B, D, E, F with the regression line and its 95% confidence interval (shading)

Figure C.12 Scatter plot for Low (triangle) and High (circle) groups for LLAMA Total with the regression line and its 95% confidence interval (shading)
C.2.4.2 Pre- and post-test correlations

As a follow-up to the paired-samples t-tests, Pearson correlations (r) between pre- and post-test scores were also calculated (see table 5.6). Correlations have been used in L2 aptitude studies to check for test-retest reliability (see Gliksman et al., 1979; Granena, 2013), where strong correlations between pre- and post-test scores (> 0.7) are considered to indicate that the post-test is measuring the same ability as the pre-test, so can be considered measure of stability in the test scores.

## pretty table

```r
knitr::kable(cor.table)
```

<table>
<thead>
<tr>
<th>LLAMA Test</th>
<th>r</th>
<th>95% CI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLAMA B</td>
<td>0.55</td>
<td>0.38-0.68</td>
<td>0.00</td>
</tr>
<tr>
<td>LLAMA D</td>
<td>0.45</td>
<td>0.26-0.61</td>
<td>0.00</td>
</tr>
<tr>
<td>LLAMA E</td>
<td>0.44</td>
<td>0.25-0.60</td>
<td>0.00</td>
</tr>
<tr>
<td>LLAMA F</td>
<td>0.27</td>
<td>0.06-0.46</td>
<td>0.01</td>
</tr>
<tr>
<td>LLAMA Total</td>
<td>0.49</td>
<td>0.30-0.63</td>
<td>0.00</td>
</tr>
</tbody>
</table>

C.3 Research Question 2: Do training factors explain more variance in post-test scores for language analytic abilities than L2 aptitude factors?

The second research question investigated the efficacy of the language analytic abilities training course in increasing LAAs, operationalised as the LLAMA F post-test scores. If the
independent variables relating to the training course explain the variance in the LLAMA F post-test scores, this would suggest that the training was responsible for the gains in test scores. However, it is also possible that any or all of the LLAMA pre-test scores are also important predictors of LLAMA F post-test scores, so these are also to be included in the model. If the training factors have higher relative importance in the model than the LLAMA pre-test scores, this would suggest that training was more predictive of increases than initial L2 aptitude levels and that training was at least partly responsible for the gains in language analytic abilities. To answer this question, a random forest regression model was fitted to discover how much of the variance in LLAMA F post-test scores can be explained by the variables of interest and also to determine which of the variables are the most predictive within the model.

A random forest regression model was fitted with the LLAMA F post-test as the response variable and with the independent variables as follows:

- pre-test scores for the LLAMA B, D, E, F, and Total
- participant approach to training in terms of total length and intensity of training
- percentage of training course completion
- participant accuracy scores over the training course
- participant time performance over the training course

C.3.1 Data preparation

Apart from the LLAMA test scores and the course completion, none of the independent variables of interest was able to be extracted or calculated directly from the raw data. The reason for this is that the training course was designed first and foremost to be an educational and learning tool, not a pure statistical measure. As a result, raw data collected from the
participants’ training had to first be prepared before being fitted to the final random forest regression model. The details of these preparations are detailed below.

C.3.1.1 Trimming times for training tasks and LLAMA tests

The time data collected from the training tasks and LLAMA pre- and post-tests contained a number of extreme outliers, with unusually large values above the median recorded. One possible explanation for this is that participants opened the training tasks and tests on their computers but were multitasking online and/or offline while also doing the training/LLAMA tests. The training/testing was then forgotten or not closed properly and so when the program window was eventually closed, a large completion time was recorded, despite the participant's not having actively done anything for most of that time.

Instead of excluding these data, completion times were trimmed by the following process. First, the data were grouped according to task or test, for example, all times from the first task of week 1 were grouped together, the times from the first task of week 2, etc. Then within each group, robust weights were calculated and times that exceeded a limit of more than 99 minutes were then reassigned a value equal to the mean, that is essentially given a weight of zero. Maintaining the extreme outliers by reassigning them a value equal to the mean allows the sample size, and measures such as variance that are dependent on the sample size, to be maintained without distorting the mean of the core of the dataset. This process retained all the data without excluding or being confounded by extreme outliers from the dataset.
C.3.1.2 Initial aptitude level

Pre-test scores were taken directly from the raw dataset and included scores for the LLAMA B, LLAMA D, LLAMA E, LLAMA F, and for the LLAMA Total. Trimmed times (see above) were also included for the total time of the combined LLAMA pre-tests. Data from participants who had not completed all the LLAMA pre-tests and the LLAMA F post-test were excluded from the analysis. The final number of participants included in the sample was, therefore, n = 90.

C.3.1.3 Approach to training course

To capture how participants chose to approach their training, a variable was created reflecting two dimensions. The first dimension captured the intensity, or frequency, with which the participants trained, operationalised as the average number of days between training sessions across the six weeks of the training course. In other words, if the participant normally massed all training tasks into one session once a week, the average number of days between sessions was seven; if the participant tended to space training sessions across all five weekdays, the average number of days between sessions was zero. The second dimension captured the total length of their training, measured in days from the first session of training to the last session. Therefore, if a participant started on the first day of the training period and finished on the last, the total duration of their training would have been 66; if they had done all the training in the last week, their total duration would have been seven. These two dimensions were then multiplied together to create a single measure of how each participant approached the training.
**C.3.1.4 Training task accuracy**

To capture the performance of participants in terms of their accuracy scores across the training course, two levels of analysis were carried out. In the first level of analysis, scores from each training task were grouped according to task type. Six task types were used in the analysis. These task types appeared more than once during the training course and are as follows: pattern matching, problem solving, grammatical judgment task, multiple choice, processing input tasks, and Google Translate tasks (see chapter 4.4.3 for detailed explanations of each task type). Missing values were omitted and then each task type was fitted to a mixed-effects linear model. The fixed effect was the mean of the cohort of participants across each task. Each task type was checked to ensure that scores for all tasks within that task type differed significantly, before continuing the analysis. That is, the data were checked to ensure that participant performance across tasks was measurably different with enough variance for the model to be informative. This fixed effect can be thought of as the average performance across the training tasks, more accurately viewed as an average profile of performance for the group. The random effect, then, was the individual variance of participants from this profile of average performance, for each task type. This resulted in a continuous measure that indicated how well each individual participant performed on average across all tasks of each task type in relation to the mean performance of all participants. Thus, the first level of analysis produced a single measure of accuracy performance for each task type to be fitted to the random forest regression model.

In the second level of analysis, the results from the random-effects model underwent a principal components analysis (Crawley 2013) to widen the scope of the analysis. The PCA resulted in two components with eigenvalues greater than 1, which are abstractions from the
mixed-effects models of accuracy scores for each training task type. If the mixed-effects models can be considered to capture the accuracy performance on the training of participants relative to the group, then the two principal components can be considered to reflect two aspects of this performance across all task types of the training. The first component is likely a reflection of ability, that is, how accurate in general a participant’s scores were for the course. The second component is less clear, but may possibly reflect the effort put into the test. However, this second more speculative component reflects another dimension in relation to which participants varied and so was considered potentially meaningful for the analysis. Consequently, the two principal components were considered as two summary variables identifying how participants varied in their accuracy scores across task types in the training course, and were fitted to the final regression models (see section 5.2.2).

C.3.1.5 Training task speed

To capture the performance of participants in terms of their time spent on each task of the training course, the same process of analysis as for the participant accuracy scores (see section 5.2.1.4) was carried out, with one exception.

Before grouping tasks by task type, missing values were summed for each participant. If the total of missing values for a participant exceeded 12, then the data from that participant were excluded from further analysis. Missing values in the remaining dataset were later imputed using an unsupervised imputation function adapted from the randomForest package in R. When a participant had more than 12 missing values in their training tasks, then the imputed values were too distant from the actual values, that is, the imputed values were unreliable and would not add predictive power to the model (see section 5.2.2). The dataset of times spent on task was then structured according to task type, using the same six task types as above. A
mixed-effect linear model was then fitted to the data for each task type to create a mean profile of training time durations for the cohort, the same process as used in 5.2.1.4 to calculate task accuracy performance. The results of the fixed effects for each task type were checked to ensure that all tasks within that task type differed significantly, before continuing the analysis. That is, participants’ duration times across tasks needed to be measurably different to create enough variance to be informative for the final models. During this process, the differences between the Google Translate tasks were not found to be significant and so were excluded from further analysis. To reiterate, for a variable to be meaningful in a model, it needs to have significant variance. If the two tests had data for a task that were not significantly different from each other, then this insignificant or insufficient variance would render redundant the inclusion of such a task. Random effects for each task type were then calculated to represent the deviation of individual participants from the cohort mean. This resulted in a continuous variable for each task, representing participants’ performance on the training course tasks in terms of time taken, to be fitted to the random forest regression model.

The random effects calculated in the first level of analysis were then subjected to a principal components analysis (Crawley 2013), again to widen the scope of the analysis. As per the training task scores, the PCA resulted in two components with eigenvalues greater than 1, which are abstractions from the mixed-effects models of accuracy scores for each training task type. If the mixed-effects models can be considered to capture the time durations performance on the training of participants relative to the group, then the two principal components can be considered to reflect two aspects of this performance across all task types of the training. The first component is likely a reflection of ability, that is, how quickly in general a participant was able to complete tasks over the whole course. The second
component is less clear, but may possibly reflect the effort put into the test. However, this second more speculative component reflects another dimension in relation to which participants varied and so was considered potentially meaningful for the analysis. Consequently, the two principal components were considered as two summary variables identifying how participants varied in their time durations across task types in the training course and were fitted to the final regression models (see section 5.2.2).

C.3.2 Random forest regression model

A random forest regression model was run with the LLAMA F post-test as the response variable and 23 potential predictor variables. The 23 variables are listed below.

LLAMA Pre-test Scores
  - LLAMA_B
  - LLAMA_D
  - LLAMA_E
  - LLAMA_F
  - LLAMA_Total

LLAMA_Time (Pre-test)

Training_Approach

Training_Completion

Training Task Accuracy
  - TTA:Pattern
  - TTA:ProblemSolving
  - TTA:GJT
- TTA:MCQ
- TTA:Processing
- TTA:Translation
- TTA:Total 1
- TTA:Total 2

Training Task Speed
- TTS:Pattern
- TTS:ProblemSolving
- TTS:GJT
- TTS:MCQ
- TTS:Processing
- TTS:Total 1
- TTS:Total 2

Table C.7 Descriptive statistics for all variables in regression models

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Range</th>
<th>S.D.</th>
<th>Variance</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLAMA_F_post</td>
<td>56.70</td>
<td>0-100</td>
<td>27.85</td>
<td>775.38</td>
<td>50.87</td>
<td>62.53</td>
</tr>
<tr>
<td>LLAMA_B</td>
<td>50.83</td>
<td>0-100</td>
<td>21.47</td>
<td>460.81</td>
<td>46.34</td>
<td>55.33</td>
</tr>
<tr>
<td>LLAMA_D</td>
<td>29.94</td>
<td>0-70</td>
<td>17.05</td>
<td>290.73</td>
<td>26.37</td>
<td>33.52</td>
</tr>
<tr>
<td>LLAMA_E</td>
<td>81.89</td>
<td>10-100</td>
<td>19.65</td>
<td>386.28</td>
<td>77.77</td>
<td>86.01</td>
</tr>
<tr>
<td>LLAMA_F</td>
<td>49.56</td>
<td>0-100</td>
<td>26.39</td>
<td>696.43</td>
<td>44.03</td>
<td>55.08</td>
</tr>
<tr>
<td>LLAMA_Total</td>
<td>212.22</td>
<td>25-330</td>
<td>51.17</td>
<td>2618.04</td>
<td>201.51</td>
<td>222.94</td>
</tr>
<tr>
<td>LLAMA_Time</td>
<td>40.75</td>
<td>1.08-88.2</td>
<td>12.11</td>
<td>146.71</td>
<td>38.22</td>
<td>43.29</td>
</tr>
<tr>
<td>Training_Approach</td>
<td>178.16</td>
<td>18-563.5</td>
<td>101.41</td>
<td>10284.17</td>
<td>156.92</td>
<td>199.40</td>
</tr>
<tr>
<td>Training_Completion</td>
<td>15.91</td>
<td>1-23</td>
<td>5.53</td>
<td>30.58</td>
<td>14.75</td>
<td>17.07</td>
</tr>
<tr>
<td>TTA_Pattern</td>
<td>0.01</td>
<td>-0.42-0.17</td>
<td>0.10</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>TTA_GJT</td>
<td>0.00</td>
<td>-0.07-0.04</td>
<td>0.03</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>
The random forest regression model was run with the following process and parameters (see chapter 4.6 for details of random forest models). Missing values were imputed to run an optimal model with 500 trees and randomly sampling 8 variables at each split. While imputing missing values does not make up for or gain data, it does allow for observed data to be retained in the dataset for analysis. Compared to discarding data which contained missing values, imputing these missing values increased the power of the analysis and reduced standard error by maintaining a larger sample size.

```
##
## Call:
## randomForest(formula = LLAMA_F_post ~ ., data = modelData_imputed)
## Type of random forest: regression
```
Table C.8 Results of random forest regression model predicting LLAMA F post-test scores

<table>
<thead>
<tr>
<th>Number of trees</th>
<th>No. of variables tried at each split</th>
<th>Mean of squared residuals</th>
<th>% Var explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>7</td>
<td>683.22</td>
<td>10.9</td>
</tr>
</tbody>
</table>
C.3.2.1 Stepwise linear regression follow up

As mentioned, the random forest regression model is an idealised and optimal model, representing the best possible analysis from the data given. As a follow-up, a stepwise linear regression model allows us to check that the most predictive variables are those given by the random forest model. To avoid inducing collinearity in the model, the LLAMA pre-test total and pre-test time duration as well as the two components for each of the training performance variables, accuracy and speed, were omitted. The predictor variables of interest used in the stepwise linear regression model are as follows:

LLAMA Pre-test Scores

- LLAMA_B
- LLAMA_D
- LLAMA_E
- LLAMA_F

Training Approach

Training Completion

Training Task Accuracy

- TTA:Pattern
- TTA:ProblemSolving
- TTA:GJT
- TTA:MCQ
- TTA:Processing
- TTA:Translation

Training Task Speed

- TTS:Pattern
• TTS:ProblemSolving
• TTS:GJT
• TTS:MCQ
• TTS:Processing
Table C.9 Pearson correlations for predictor variables (listed) with response variable of LLAMA F post-test scores

<table>
<thead>
<tr>
<th>Variables</th>
<th>r</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLAMA_F_post</td>
<td>1.00</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>LLAMA_B</td>
<td>0.24</td>
<td>0.026</td>
</tr>
<tr>
<td>LLAMA_D</td>
<td>-0.19</td>
<td>0.078</td>
</tr>
<tr>
<td>LLAMA_E</td>
<td>0.36</td>
<td>0.001</td>
</tr>
<tr>
<td>LLAMA_F</td>
<td>0.27</td>
<td>0.009</td>
</tr>
<tr>
<td>Training_Appearance</td>
<td>-0.11</td>
<td>0.289</td>
</tr>
<tr>
<td>Training_Completion</td>
<td>-0.04</td>
<td>0.728</td>
</tr>
<tr>
<td>TTA_Pattern</td>
<td>0.23</td>
<td>0.03</td>
</tr>
<tr>
<td>TTA_GJT</td>
<td>0.06</td>
<td>0.576</td>
</tr>
<tr>
<td>TTA_MCQ</td>
<td>0.10</td>
<td>0.339</td>
</tr>
<tr>
<td>TTA_Processing</td>
<td>0.30</td>
<td>0.004</td>
</tr>
<tr>
<td>TTA_ProblemSolving</td>
<td>0.24</td>
<td>0.025</td>
</tr>
<tr>
<td>TTA_Translation</td>
<td>0.14</td>
<td>0.173</td>
</tr>
<tr>
<td>TTS_Pattern</td>
<td>0.17</td>
<td>0.119</td>
</tr>
<tr>
<td>TTS_GJT</td>
<td>0.11</td>
<td>0.319</td>
</tr>
<tr>
<td>TTS_MCQ</td>
<td>0.15</td>
<td>0.165</td>
</tr>
<tr>
<td>TTS_Processing</td>
<td>0.04</td>
<td>0.709</td>
</tr>
<tr>
<td>TTS_ProblemSolving</td>
<td>0.07</td>
<td>0.503</td>
</tr>
</tbody>
</table>
## Single term deletions

## Model:

```R
LLAMA_F_post ~ LLAMA_B + LLAMA_D + LLAMA_E + LLAMA_F + 
Training_Approach + 
  Training_Completion + TTA_Pattern + TTA_GJT + TTA_MCQ + 
  TTA_Processing + 
  TTA_ProblemSolving + TTA_Translation + TTS_Pattern + TTS_GJT + 
  TTS_MCQ + TTS_Processing + TTS_ProblemSolving
```

##

<table>
<thead>
<tr>
<th>Term</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
<th>F Value</th>
<th>Pr(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTA_Translation</td>
<td>1</td>
<td>0.09</td>
<td>43606 590.48</td>
<td>0.0001</td>
<td>0.99047</td>
<td></td>
</tr>
<tr>
<td>TTA_MCQ</td>
<td>1</td>
<td>0.79</td>
<td>43607 590.48</td>
<td>0.0013</td>
<td>0.97133</td>
<td></td>
</tr>
<tr>
<td>TTS_ProblemSolving</td>
<td>1</td>
<td>32.60</td>
<td>43639 590.55</td>
<td>0.0538</td>
<td>0.81720</td>
<td></td>
</tr>
<tr>
<td>TTS_MCQ</td>
<td>1</td>
<td>35.14</td>
<td>43641 590.56</td>
<td>0.0580</td>
<td>0.81034</td>
<td></td>
</tr>
<tr>
<td>TTS_GJT</td>
<td>1</td>
<td>44.93</td>
<td>43651 590.58</td>
<td>0.0742</td>
<td>0.78611</td>
<td></td>
</tr>
<tr>
<td>TTA_GJT</td>
<td>1</td>
<td>179.56</td>
<td>43786 590.85</td>
<td>0.2965</td>
<td>0.58778</td>
<td></td>
</tr>
<tr>
<td>Training_Approach</td>
<td>1</td>
<td>344.62</td>
<td>43951 591.19</td>
<td>0.5690</td>
<td>0.45311</td>
<td></td>
</tr>
<tr>
<td>LLAMA_B</td>
<td>1</td>
<td>488.15</td>
<td>44094 591.48</td>
<td>0.8060</td>
<td>0.37230</td>
<td></td>
</tr>
<tr>
<td>TTA_ProblemSolving</td>
<td>1</td>
<td>531.29</td>
<td>44138 591.57</td>
<td>0.8772</td>
<td>0.35209</td>
<td></td>
</tr>
<tr>
<td>Training_Completion</td>
<td>1</td>
<td>618.95</td>
<td>44225 591.75</td>
<td>1.0220</td>
<td>0.31544</td>
<td></td>
</tr>
<tr>
<td>TTA_Pattern</td>
<td>1</td>
<td>978.83</td>
<td>44585 592.48</td>
<td>1.6162</td>
<td>0.20772</td>
<td></td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>43606</td>
<td>592.48 592.48</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## LLAMA_F
  1 1020.13 44626 592.56 1.6844 0.19849

## TTA_Processing
  1 1479.01 45085 593.49 2.4421 0.12251

## LLAMA_D
  1 2130.46 45737 594.78 3.5177 0.06477

## TTS_Processing
  1 2256.97 45863 595.02 3.7266 0.05749

## LLAMA_E
  1 2401.03 46007 595.31 3.9644 0.05027

## TTS_Pattern
  1 2988.93 46595 596.45 4.9351 0.02946 *

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Start:  AIC=592.48

## LLAMA_F_post ~ LLAMA_B + LLAMA_D + LLAMA_E + LLAMA_F +
   Training_Approach +
   Training_Completion + TTA_Pattern + TTA_GJT + TTA_MCQ +
   TTA_Processing +
   TTA_ProblemSolving + TTA_Translation + TTS_Pattern + TTS_GJT +
   TTS_MCQ + TTS_Processing + TTS_ProblemSolving

## Df  Sum of Sq   RSS    AIC
## - TTA_Translation  1   0.09 43606 590.48
## - TTA_MCQ          1   0.79 43607 590.48
## - TTS_ProblemSolving  1  32.60 43639 590.55
## - TTS_MCQ          1   35.14 43641 590.56
## - TTS_GJT          1   44.93 43651 590.58
## - TTA_GJT          1  179.56 43786 590.85
<table>
<thead>
<tr>
<th>Variable</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training_Approach</td>
<td>1</td>
<td>344.62</td>
<td>43951</td>
<td>591.19</td>
</tr>
<tr>
<td>LLAMA_B</td>
<td>1</td>
<td>488.15</td>
<td>44094</td>
<td>591.48</td>
</tr>
<tr>
<td>TTA_ProblemSolving</td>
<td>1</td>
<td>531.29</td>
<td>44138</td>
<td>591.57</td>
</tr>
<tr>
<td>Training_Completion</td>
<td>1</td>
<td>618.95</td>
<td>44225</td>
<td>591.75</td>
</tr>
<tr>
<td>TTA_Pattern</td>
<td>1</td>
<td>978.83</td>
<td>44585</td>
<td>592.48</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td></td>
<td>43606</td>
<td>592.48</td>
</tr>
<tr>
<td>LLAMA_F</td>
<td>1</td>
<td>1020.13</td>
<td>44626</td>
<td>592.56</td>
</tr>
<tr>
<td>TTA_Processing</td>
<td>1</td>
<td>1479.01</td>
<td>45085</td>
<td>593.49</td>
</tr>
<tr>
<td>LLAMA_D</td>
<td>1</td>
<td>2130.46</td>
<td>45737</td>
<td>594.78</td>
</tr>
<tr>
<td>TTS_Processing</td>
<td>1</td>
<td>2256.97</td>
<td>45863</td>
<td>595.02</td>
</tr>
<tr>
<td>LLAMA_E</td>
<td>1</td>
<td>2401.03</td>
<td>46007</td>
<td>595.31</td>
</tr>
<tr>
<td>TTS_Pattern</td>
<td>1</td>
<td>2988.93</td>
<td>46595</td>
<td>596.45</td>
</tr>
</tbody>
</table>

## Step: AIC=590.48

\[
\text{LLAMA_F} \sim \text{LLAMA_B} + \text{LLAMA_D} + \text{LLAMA_E} + \text{LLAMA_F} + \text{Training_Approach} + \text{Training_Completion} + \text{TTA_Pattern} + \text{TTA_GJT} + \text{TTA_MCQ} + \text{TTA_Processing} + \text{TTA_ProblemSolving} + \text{TTS_Pattern} + \text{TTS_GJT} + \text{TTS_MCQ} + \text{TTS_Processing} + \text{TTS_ProblemSolving}.
\]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTA_MCQ</td>
<td>1</td>
<td>0.82</td>
<td>43607</td>
<td>588.48</td>
</tr>
</tbody>
</table>
### Step: AIC=588.48

```
# LLAMA_F_post ~ LLAMA_B + LLAMA_D + LLAMA_E + LLAMA_F +
# Training_Approach +
# Training_Completion + TTA_Pattern + TTA_GJT + TTA_Processing +
# TTA_ProblemSolving + TTS_Pattern + TTS_GJT + TTS_MCQ +
# TTS_Processing +
```
## TTS_ProblemSolving

<table>
<thead>
<tr>
<th>Model</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>- TTS_ProblemSolving</td>
<td>1</td>
<td>32.14</td>
<td>43639</td>
<td>586.55</td>
</tr>
<tr>
<td>- TTS_MCQ</td>
<td>1</td>
<td>39.73</td>
<td>43647</td>
<td>586.57</td>
</tr>
<tr>
<td>- TTS_GJT</td>
<td>1</td>
<td>47.86</td>
<td>43655</td>
<td>586.58</td>
</tr>
<tr>
<td>- TTA_GJT</td>
<td>1</td>
<td>225.80</td>
<td>43833</td>
<td>586.95</td>
</tr>
<tr>
<td>- Training_Approach</td>
<td>1</td>
<td>347.20</td>
<td>43954</td>
<td>587.20</td>
</tr>
<tr>
<td>- LLAMA_B</td>
<td>1</td>
<td>491.89</td>
<td>44099</td>
<td>587.49</td>
</tr>
<tr>
<td>- TTA_ProblemSolving</td>
<td>1</td>
<td>540.59</td>
<td>44148</td>
<td>587.59</td>
</tr>
<tr>
<td>- Training_Completion</td>
<td>1</td>
<td>624.81</td>
<td>44232</td>
<td>587.77</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td></td>
<td>43607</td>
<td>588.48</td>
</tr>
<tr>
<td>- TTA_Pattern</td>
<td>1</td>
<td>1053.22</td>
<td>44660</td>
<td>588.63</td>
</tr>
<tr>
<td>- LLAMA_F</td>
<td>1</td>
<td>1118.13</td>
<td>44725</td>
<td>588.76</td>
</tr>
<tr>
<td>- TTA_Processing</td>
<td>1</td>
<td>1508.25</td>
<td>45115</td>
<td>589.55</td>
</tr>
<tr>
<td>- LLAMA_D</td>
<td>1</td>
<td>2133.87</td>
<td>45741</td>
<td>590.78</td>
</tr>
<tr>
<td>- TTS_Processing</td>
<td>1</td>
<td>2260.26</td>
<td>45867</td>
<td>591.03</td>
</tr>
<tr>
<td>- LLAMA_E</td>
<td>1</td>
<td>2646.97</td>
<td>46254</td>
<td>591.79</td>
</tr>
<tr>
<td>- TTS_Pattern</td>
<td>1</td>
<td>2992.80</td>
<td>46600</td>
<td>592.46</td>
</tr>
</tbody>
</table>

## Step:  AIC=586.55

```
## LLAMA_F_post ~ LLAMA_B + LLAMA_D + LLAMA_E + LLAMA_F +
Training_Approach +
## Training_Completion + TTA_Pattern + TTA_GJT + TTA_Processing
```
## TTA_ProblemSolving + TTS_Pattern + TTS_GJT + TTS_MCQ + TTS_Processing

##

<table>
<thead>
<tr>
<th>Model</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>- TTS_MCQ</td>
<td>1</td>
<td>54.67</td>
<td>43694</td>
<td>584.66</td>
</tr>
<tr>
<td>- TTS_GJT</td>
<td>1</td>
<td>70.32</td>
<td>43710</td>
<td>584.70</td>
</tr>
<tr>
<td>- TTA_GJT</td>
<td>1</td>
<td>264.52</td>
<td>43904</td>
<td>585.10</td>
</tr>
<tr>
<td>- Training_Approach</td>
<td>1</td>
<td>381.14</td>
<td>44020</td>
<td>585.33</td>
</tr>
<tr>
<td>- LLAMA_B</td>
<td>1</td>
<td>476.43</td>
<td>44116</td>
<td>585.53</td>
</tr>
<tr>
<td>- Training_Completion</td>
<td>1</td>
<td>604.97</td>
<td>44244</td>
<td>585.79</td>
</tr>
<tr>
<td>- TTA_ProblemSolving</td>
<td>1</td>
<td>785.99</td>
<td>44425</td>
<td>586.16</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td></td>
<td>43639</td>
<td>586.55</td>
</tr>
<tr>
<td>- TTA_Pattern</td>
<td>1</td>
<td>1022.73</td>
<td>44662</td>
<td>586.64</td>
</tr>
<tr>
<td>- LLAMA_F</td>
<td>1</td>
<td>1087.28</td>
<td>44727</td>
<td>586.77</td>
</tr>
<tr>
<td>- TTA_Processing</td>
<td>1</td>
<td>1483.23</td>
<td>45122</td>
<td>587.56</td>
</tr>
<tr>
<td>- TTS_Processing</td>
<td>1</td>
<td>2280.11</td>
<td>45919</td>
<td>589.13</td>
</tr>
<tr>
<td>- LLAMA_D</td>
<td>1</td>
<td>2384.23</td>
<td>46023</td>
<td>589.34</td>
</tr>
<tr>
<td>- LLAMA_E</td>
<td>1</td>
<td>2655.31</td>
<td>46295</td>
<td>589.87</td>
</tr>
<tr>
<td>- TTS_Pattern</td>
<td>1</td>
<td>3059.58</td>
<td>46699</td>
<td>590.65</td>
</tr>
</tbody>
</table>

##

## Step:  AIC=584.66

## LLAMA_F_post ~ LLAMA_B + LLAMA_D + LLAMA_E + LLAMA_F + Training_Approach +

346
## Training_Completion + TTA_Pattern + TTA_GJT + TTA_Processing + TTA_ProblemSolving + TTS_Pattern + TTS_GJT + TTS_Processing

##

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>- TTS_GJT</td>
<td>1</td>
<td>48.7</td>
<td>43743</td>
<td>582.76</td>
</tr>
<tr>
<td>- TTA_GJT</td>
<td>1</td>
<td>261.9</td>
<td>43956</td>
<td>583.20</td>
</tr>
<tr>
<td>- Training_Approach</td>
<td>1</td>
<td>328.0</td>
<td>44022</td>
<td>583.34</td>
</tr>
<tr>
<td>- LLAMA_B</td>
<td>1</td>
<td>530.1</td>
<td>44224</td>
<td>583.75</td>
</tr>
<tr>
<td>- Training_Completion</td>
<td>1</td>
<td>657.3</td>
<td>44351</td>
<td>584.01</td>
</tr>
<tr>
<td>- TTA_ProblemSolving</td>
<td>1</td>
<td>799.8</td>
<td>44494</td>
<td>584.30</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td></td>
<td>43694</td>
<td>584.66</td>
</tr>
<tr>
<td>- LLAMA_F</td>
<td>1</td>
<td>1177.8</td>
<td>44872</td>
<td>585.06</td>
</tr>
<tr>
<td>- TTA_Pattern</td>
<td>1</td>
<td>1226.8</td>
<td>44921</td>
<td>585.16</td>
</tr>
<tr>
<td>- TTA_Processing</td>
<td>1</td>
<td>1457.4</td>
<td>45151</td>
<td>585.62</td>
</tr>
<tr>
<td>- TTS_Processing</td>
<td>1</td>
<td>2225.7</td>
<td>45920</td>
<td>587.14</td>
</tr>
<tr>
<td>- LLAMA_D</td>
<td>1</td>
<td>2336.5</td>
<td>46030</td>
<td>587.35</td>
</tr>
<tr>
<td>- LLAMA_E</td>
<td>1</td>
<td>2626.3</td>
<td>46320</td>
<td>587.92</td>
</tr>
<tr>
<td>- TTS_Pattern</td>
<td>1</td>
<td>4072.1</td>
<td>47766</td>
<td>590.68</td>
</tr>
</tbody>
</table>

##

## Step: AIC=582.76

## LLAMA_F_post ~ LLAMA_B + LLAMA_D + LLAMA_E + LLAMA_F + Training_Approach + Training_Completion + TTA_Pattern + TTA_GJT + TTA_Processing
## TTA_ProblemSolving + TTS_Pattern + TTS_Processing

##

<table>
<thead>
<tr>
<th>Model</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>- TTA_GJT</td>
<td>1</td>
<td>238.3</td>
<td>43981</td>
<td>581.25</td>
</tr>
<tr>
<td>- Training_Approach</td>
<td>1</td>
<td>352.4</td>
<td>44095</td>
<td>581.49</td>
</tr>
<tr>
<td>- LLAMA_B</td>
<td>1</td>
<td>524.7</td>
<td>44267</td>
<td>581.84</td>
</tr>
<tr>
<td>- Training_Completion</td>
<td>1</td>
<td>643.6</td>
<td>44386</td>
<td>582.08</td>
</tr>
<tr>
<td>- TTA_ProblemSolving</td>
<td>1</td>
<td>756.0</td>
<td>44499</td>
<td>582.31</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td></td>
<td>43743</td>
<td>582.76</td>
</tr>
<tr>
<td>- LLAMA_F</td>
<td>1</td>
<td>1132.3</td>
<td>44875</td>
<td>583.06</td>
</tr>
<tr>
<td>- TTA_Pattern</td>
<td>1</td>
<td>1345.8</td>
<td>45088</td>
<td>583.49</td>
</tr>
<tr>
<td>- TTA_Processing</td>
<td>1</td>
<td>1540.0</td>
<td>45283</td>
<td>583.88</td>
</tr>
<tr>
<td>- LLAMA_D</td>
<td>1</td>
<td>2299.7</td>
<td>46042</td>
<td>585.38</td>
</tr>
<tr>
<td>- LLAMA_E</td>
<td>1</td>
<td>2578.2</td>
<td>46321</td>
<td>585.92</td>
</tr>
<tr>
<td>- TTS_Processing</td>
<td>1</td>
<td>2593.8</td>
<td>46336</td>
<td>585.95</td>
</tr>
<tr>
<td>- TTS_Pattern</td>
<td>1</td>
<td>4023.5</td>
<td>47766</td>
<td>588.68</td>
</tr>
</tbody>
</table>

## Step:  AIC=581.25

## LLAMA_F_post ~ LLAMA_B + LLAMA_D + LLAMA_E + LLAMA_F + Training_Approach + Training_Completion + TTA_Pattern + TTA_Processing + TTA_ProblemSolving + TTS_Pattern + TTS_Processing
##
## | Df | Sum of Sq | RSS | AIC |
## |----|-----------|-----|-----|
## |  1 |   246.5   | 44228 | 579.76 |
## |  1 |   481.8   | 44463 | 580.23 |
## |  1 |   654.9   | 44636 | 580.58 |
## |  1 |   752.2   | 44733 | 580.78 |
## |    |           | 43981 | 581.25 |
## |  1 |  1210.5   | 45192 | 581.70 |
## |  1 |  1373.0   | 45354 | 582.02 |
## |  1 |  1457.9   | 45439 | 582.19 |
## |  1 |  2395.9   | 46377 | 584.03 |
## |  1 |  2755.5   | 46736 | 584.72 |
## |  1 |  2965.7   | 46947 | 585.13 |
## |  1 |  4080.4   | 48061 | 587.24 |
##
## Step: AIC=579.76

## LLAMA_F_post ~ LLAMA_B + LLAMA_D + LLAMA_E + LLAMA_F + Training_Completion +
## TTA_Pattern + TTA_Processing + TTA_ProblemSolving +
## TTS_Pattern +
## TTS_Processing

##
## | Df | Sum of Sq | RSS | AIC |
## |----|-----------|-----|-----|
## |  1 |   571.4   | 44299 | 578.91 |
## - TTA_ProblemSolving  1    767.8 44995 579.31
## <none>                             44228 579.76
## - Training_Completion  1    1260.3 45488 580.29
## - TTA_Pattern          1    1285.8 45513 580.34
## - TTA_Processing       1    1409.2 45637 580.58
## - LLAMA_F              1    1871.0 46099 581.49
## - LLAMA_E              1    2208.8 46436 582.14
## - TTS_Processing       1    2757.7 46985 583.20
## - LLAMA_D              1    2978.1 47206 583.62
## - TTS_Pattern          1    3862.1 48090 585.29
##
## Step:  AIC=578.91
## LLAMA_F_post ~ LLAMA_D + LLAMA_E + LLAMA_F + Training_Completion
## +
##     TTA_Pattern + TTA_Processing + TTA_ProblemSolving +
## TTS_Pattern +
##
##                       Df Sum of Sq   RSS    AIC
## - TTA_ProblemSolving   1     802.9 45602 578.51
## <none>                             44799 578.91
## - TTA_Pattern          1    1157.4 45956 579.21
## - Training_Completion  1    1252.1 46051 579.39
## - TTA_Processing       1    1661.7 46461 580.19
## - LLAMA_F              1    2271.2 47070 581.36
## - LLAMA_D              1    2768.0 47567 582.31
## - LLAMA_E              1    3074.9 47874 582.89
## - TTS_Processing       1    3488.7 48288 583.66
## - TTS_Pattern          1    3933.6 48732 584.49
##
## Step:  AIC=578.51
## LLAMA_F_post ~ LLAMA_D + LLAMA_E + LLAMA_F + Training_Completion
##      +
##      TTA_Pattern + TTA_Processing + TTS_Pattern + TTS_Processing
##
##                       Df Sum of Sq   RSS    AIC
## <none>                             45602 578.51
## - Training_Completion  1    1146.1 46748 578.74
## - TTA_Pattern          1    1627.9 47230 579.67
## - LLAMA_D              1    2257.4 47859 580.86
## - LLAMA_F              1    2589.2 48191 581.48
## - LLAMA_E              1    2952.0 48554 582.16
## - TTS_Processing       1    3239.9 48842 582.69
## - TTA_Processing       1    3678.2 49280 583.49
## - TTS_Pattern          1    3744.5 49346 583.61
##
## Single term deletions
##
## Model:
## LLAMA_F_post ~ LLAMA_D + LLAMA_E + LLAMA_F + Training_Completion + 
## TTA_Pattern + TTA_Processing + TTS_Pattern + TTS_Processing 
##
## Df Sum of Sq    RSS    AIC  F Value    Pr(F) 
## <none>                   45602 578.51 
## Training_Completion  1 1146.1 46748 578.74  2.0358 0.15748 
## TTA_Pattern          1 1627.9 47230 579.67  2.8915 0.09288 . 
## LLAMA_D              1 2257.4 47859 580.86  4.0096 0.04859 * 
## LLAMA_F              1 2589.2 48191 581.48  4.5990 0.03499 * 
## LLAMA_E              1 2952.0 48554 582.16  5.2435 0.02463 * 
## TTS_Processing       1 3239.9 48842 582.69  5.7548 0.01874 * 
## TTA_Processing       1 3678.2 49280 583.49  6.5333 0.01246 * 
## TTS_Pattern          1 3744.5 49346 583.61  6.6512 0.01172 * 
##
## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

##
## Call: 
## lm(formula = LLAMA_F_post ~ LLAMA_E + LLAMA_F + TTA_Processing + 
## TTS_Pattern + TTS_Processing, data = lmData)
##
## Residuals: 
## Min  1Q Median  3Q Max 
## -68.192 -15.023  6.075  16.748  37.105 
##
## Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 18.1959  | 11.9159    | 1.527   | 0.13051  |
| LLAMA_E        | 0.3283   | 0.1467     | 2.237   | 0.02793  *|
| LLAMA_F        | 0.2368   | 0.1032     | 2.295   | 0.02423  *|
| TTA_Processing | 155.5734 | 54.5108    | 2.854   | 0.00544  **|
| TTS_Pattern    | 4.6424   | 2.0961     | 2.215   | 0.02948  *|
| TTS_Processing | -7.4747  | 2.9792     | -2.509  | 0.01403  *|

---

## Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 . ‘.’ 0.1 ‘ ’ 1

## Residual standard error: 24.5 on 84 degrees of freedom

## Multiple R-squared: 0.2693, Adjusted R-squared: 0.2258

## F-statistic: 6.193 on 5 and 84 DF,  p-value: 6.201e-05
<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SEB</th>
<th>β</th>
<th>Significance (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>18.1959</td>
<td>11.9159</td>
<td>1.527</td>
<td>0.13051</td>
</tr>
<tr>
<td>LLAMA_E</td>
<td>0.3283</td>
<td>0.1467</td>
<td>2.237</td>
<td>0.02793 *</td>
</tr>
<tr>
<td>LLAMA_F</td>
<td>0.2368</td>
<td>0.1032</td>
<td>2.295</td>
<td>0.02423 *</td>
</tr>
<tr>
<td>TTA:Processing</td>
<td>155.5734</td>
<td>54.5108</td>
<td>2.854</td>
<td>0.00544 **</td>
</tr>
<tr>
<td>TTS:Pattern</td>
<td>4.6424</td>
<td>2.0961</td>
<td>2.215</td>
<td>0.02948 *</td>
</tr>
<tr>
<td>TTS:Processing</td>
<td>-7.4747</td>
<td>2.9792</td>
<td>-2.509</td>
<td>0.01403 *</td>
</tr>
<tr>
<td>Multiple $R^2$</td>
<td>0.2693</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.2258</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F(5,84)$</td>
<td>6.193</td>
<td></td>
<td></td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
REFERENCES


http://lltf.net/mlat-sample-items

http://lltf.net/aptitude-tests/language-aptitude-tests/pimsleur-language-aptitude-battery/plab-sample-items/


VanPatten, B. (2010). Some verbs are more perfect than others: Why learners have difficulty with ser and estar and what it means for instruction. *Hispania, 93*(1), 29–38.


*Applied Psycholinguistics, 3*(2), 127–140.


