Explaining differences in adult second language learning
The role of language input characteristics and learners’ cognitive aptitudes
Ćurčić, M.

Citation for published version (APA):
Maja Ćurčić

Explaining differences in adult second language learning: The role of language input characteristics and learners’ cognitive aptitudes

This dissertation examines how adult learning of novel language structures is affected by the characteristics of the language input that learners are exposed to and by learners’ cognitive aptitudes, such as analytical ability and working memory capacity.

In a series of experiments, adult native speakers of Dutch received brief auditory exposure to a miniature language based on Fijian. Within the language, learners were also exposed to an agreement pattern between determiners and nouns: masculine determiner *lep* preceded nouns ending in –uk, and feminine determiner *ris* preceded nouns ending in –is (e.g., *lep oseuk*, *ris burogis*). After the exposure, learners’ knowledge of the agreement pattern and their ability to quickly predict nouns based on determiners were measured using grammaticality judgment tasks, a production task, and a processing task with eye-tracking.

There were substantial differences between learners, both in terms of their learning success and the awareness of the underlying language patterns. The reliability of the language input (i.e., whether the agreement pattern in the input was fully regular or featured exceptions) affected learning in different ways depending on the length of the language exposure and learners’ awareness of the agreement structure. Learners’ working memory capacity and analytical abilities were positively related to their learning outcomes and to the occurrence of pattern awareness among learners. Some learners developed quick predictive processing based on determiners, but this was always accompanied by learners’ awareness that determiners were useful for making choices during the test with eye-tracking.
Explaining differences in adult second language learning: The role of language input characteristics and learners’ cognitive aptitudes
Explaining differences in adult second language learning: The role of language input characteristics and learners’ cognitive aptitudes

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College van Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 1 februari 2018, te 12.00 uur

door

Maja Ćurčić

geboren te Novi Sad, Joegoslavië
Promotiecommissie:

**Promotor:** Prof. dr. F. Kuiken
Universiteit van Amsterdam

**Copromotor:** Dr. S.J. Andringa
Universiteit van Amsterdam

**Overige leden:**
- Prof. dr. E.O. Aboh
  Universiteit van Amsterdam
- Prof. dr. P.E. Dussias
  Penn State University
- Prof. dr. J.H. Hulstijn
  Universiteit van Amsterdam
- Prof. dr. W.M. Lowie
  Rijksuniversiteit Groningen
- Dr. E. Marsden
  University of York
- Dr. J.E. Rispens
  Universiteit van Amsterdam

Faculteit der Geesteswetenschappen
# Table of Contents

Acknowledgements ........................................................................................................... v

List of abbreviations ......................................................................................................... ix

Author contributions ......................................................................................................... xi

Chapter 1: Introduction ....................................................................................................... 1
  1.1 Input characteristics and language acquisition ......................................................... 2
  1.2 Cognitive aptitudes and second language acquisition.................................................. 4
  1.3 Individual differences in awareness during exposure .................................................... 6
  1.4 Individual differences in L2 processing ......................................................................... 8
  1.5 Goals and research questions ...................................................................................... 9
  1.6 Outline of the dissertation ........................................................................................... 10

Chapter 2: The role of unreliable input in L2 pattern learning .......................................... 13
  Abstract ............................................................................................................................. 13
  2.1 Introduction .................................................................................................................. 13
    2.1.1 Input characteristics and L2 pattern learning ......................................................... 14
    2.1.2 Aptitudes and L2 pattern learning ......................................................................... 16
    2.1.3 The present study .................................................................................................. 18
  2.2 Method .......................................................................................................................... 19
    2.2.1 Participants ............................................................................................................ 19
    2.2.2 The target language and the target structure ......................................................... 20
    2.2.3 Language exposure ............................................................................................... 21
    2.2.4 Oral GJTs ............................................................................................................... 24
    2.2.5 Debriefing ............................................................................................................ 25
    2.2.6 Aptitude measures ............................................................................................... 25
    2.2.7 Statistical procedures ........................................................................................... 27
  2.3 Results ......................................................................................................................... 28
    2.3.1 Debriefing results – awareness .............................................................................. 28
    2.3.2 GJT with familiar nouns ....................................................................................... 29
    2.3.3 GJT with novel nouns .......................................................................................... 35
    2.3.4 The role of working memory, rote memory, analytical ability, and statistical learning ability .................................................................................................................. 37
  2.4 Discussion ..................................................................................................................... 40
Chapter 3: Effects of prolonged exposure on L2 pattern learning from reliable and unreliable input .............................................. 45

Abstract ...................................................................................... 45

3.1 Introduction ............................................................................. 45

3.1.1 Role of input reliability in L2 pattern learning .......... 46

3.1.2 Implicit learning and awareness – how they interact and develop over time ........................................ 47

3.1.3 The present study .............................................................. 49

3.2 Method ..................................................................................... 50

3.2.1 Participants .......................................................................... 50

3.2.2 The target language and exposure .................................. 50

3.2.3 Oral GJTs ............................................................................. 51

3.2.4 Debriefing ............................................................................ 51

3.2.5 Aptitude measures ............................................................... 51

3.2.6 Statistical procedures ........................................................ 53

3.3 Results – replication of the initial study ......................... 54

3.3.1 Debriefing results – awareness ........................................ 54

3.3.2 GJT with familiar nouns .................................................... 54

3.3.3 GJT with novel nouns ......................................................... 60

3.3.4 The role of working memory, rote memory, and analytical ability ......................................................... 62

3.4 Results – effects of longer exposure on implicit learning and awareness ............................................................. 67

3.5 Discussion ................................................................................. 70

Chapter 4: Effects of input rehearsal on L2 pattern learning .......... 75

Abstract ...................................................................................... 75

4.1 Introduction ............................................................................. 75

4.1.1 Producing language and its role in SLA ......................... 76

4.1.2 Rehearsal effects as a neglected but promising area of research .......................................................... 77

4.1.3 The role of rehearsal in short-term memory ............... 77

4.1.4 Effects of rehearsal on vocabulary and pattern learning ............................................................. 78

4.1.5 The present study .............................................................. 80

4.2 Method ..................................................................................... 81

4.2.1 Participants .......................................................................... 81

4.2.2 The target language and the target structure .......... 82

4.2.3 Language exposure ............................................................. 82

4.2.4 Production test ................................................................. 85

4.2.5 Oral GJTs ............................................................................. 85
4.2.6 Debriefing ................................................................. 86
4.2.7 Aptitude measures .................................................. 86
4.2.8 Statistical procedures .............................................. 88

4.3 Results ........................................................................... 89
4.3.1 Debriefing results – awareness .................................. 89
4.3.2 GJT with familiar nouns ........................................... 90
4.3.3 GJT with novel nouns ............................................... 93
4.3.4 Production test ......................................................... 95
4.3.5 The role of working memory, rote memory, and analytical ability ........................................... 98

4.4 Discussion ...................................................................... 102

Chapter 5: The role of awareness, cognitive aptitudes, and instruction type in L2 processing of determiners ................................................. 107

Abstract ............................................................................. 107

5.1 Introduction ..................................................................... 107
5.1.1 Language processing in L1 speakers ....................... 108
5.1.2 Language processing in L2 learners ....................... 109
5.1.3 Factors contributing to L2 processing ................... 110
5.1.4 The present studies .................................................. 114

5.2 Method – Study 1 ......................................................... 115
5.2.1 Participants .............................................................. 115
5.2.2 The target language and the target structure ....... 115
5.2.3 Language exposure ................................................... 116
5.2.4 The test with eye-tracking ...................................... 119
5.2.5 Debriefing .............................................................. 122
5.2.6 Aptitude measures ................................................... 122

5.3 Results – Study 1 ......................................................... 123
5.3.1 Debriefing results – awareness .............................. 123
5.3.2 Eye movement results ............................................. 124
5.3.3 Reaction times (RTs) ............................................... 130

5.4 Method – Study 2 ......................................................... 132
5.4.1 Participants .............................................................. 132
5.4.2 Language exposure ................................................... 132
5.4.3 The test with eye-tracking ...................................... 133
5.4.4 Debriefing .............................................................. 133
5.4.5 Aptitude measures ................................................... 133

5.5 Results – Study 2 ......................................................... 134
5.5.1 Debriefing results – awareness .............................. 134
5.5.2 Eye movement results ............................................. 135
5.5.3 Reaction times (RTs) ............................................... 137
# Table of contents

5.6 Discussion ........................................................................................................ 139

Chapter 6: General discussion and conclusions .................................................. 143
   6.1 How reliability of the input affected language learning .......................... 143
   6.2 Aptitudes and pattern learning ................................................................. 145
   6.3 Individual differences in L2 processing .................................................... 149
   6.4 Awareness-related differences ................................................................. 151
   6.5 Methodological considerations ............................................................... 152
   6.6 Practical implications .............................................................................. 155

Appendices .............................................................................................................. 157
   Appendix A: Target language words used in the exposure .......................... 157
   Appendix B: Target structure exposure items ............................................. 159
   Appendix C: GJT with novel nouns ............................................................... 165
   Appendix D: GJT with familiar nouns ......................................................... 166
   Appendix E: Test with eye-tracking ............................................................... 167

References .............................................................................................................. 169

Summary .................................................................................................................. 183

Samenvatting ........................................................................................................... 189

Curriculum vitae .................................................................................................... 197
Acknowledgements

I am now at the end of my four-year PhD research journey. At the beginning of this journey, I saw PhD as an exam where knowledge, intellect, and hard work are tested. However, I gradually realized that it is an exam of mental and psychological as much as intellectual strength, and it is not only an exam that tests you but also an exam that makes you stronger! This section will probably be too short to paint an accurate picture of all my PhD experiences and feelings, and express my gratitude to the many people that I feel indebted to, but this time I will step away from the PhD trait of perfectionism, and simply apologize if I forgot to mention someone.

First of all, I am endlessly grateful to my supervisors Sible Andringa and Folkert Kuiken. Sible, where shall I start in expressing my gratitude? You supervised my Master thesis, and at that time I already knew it would be difficult to find a better supervisor than you for my PhD. Your valuable insights and sharp, detailed feedback made me improve my research and writing skills every day. I owe you a lot for your endless patience and availability for random, out of the blue meetings. They meant so much to me! Also, thank you for your endless energy, enthusiasm, and encouragement, which were very important sources of motivation for me. Folkert, I am grateful to you for so many things. You were an excellent supervisor who supervised the project and its progress professionally and pragmatically at the same time, which taught me a lot about being optimal and pragmatic, the qualities needed to bring a PhD project to an end. “There is also a life after a PhD” and “You should also keep something for a postdoc” were your favorite sentences to encourage me to think in terms of what is feasible. Also, thank you for being so kind to me, making it possible for me to follow Dutch courses, and last, but not least for always speaking Dutch to me.

A number of people were directly or indirectly involved in my project and helped me at different stages. First, I wish to express my endless gratitude to Dirk Vet, the most kind-hearted lab technician in the world. Dirk, thank you for your invaluable and generous technical assistance during my experiments, and also for the many inspiring chats we had about academia and different world issues. Working and chatting with you was a real pleasure. I am very grateful to my dear friend Ana Bosnić for being so kind and brave to become the first and only native speaker of my invented language, for recording the auditory materials for my experiments, and for providing an incredible role model for my learners. Ana, I spent over 300 hours in the lab listening to your great voice, and I enjoyed it! I wish to thank Eva van Lier who was so kind to record
instructions for my experiments and to help me in creating the language. I am also grateful to Iris Broedelet for her time, effort, and attention to detail while helping me analyze the results of the production and non-word repetition tasks. I feel indebted to more than 150 participants of my experiment who took their time to come to the lab and did their best to learn my non-existing language. Finally, I am grateful to my committee members for being the persons who took the time and effort to read my dissertation in its entirety and assess it: Emma Marsden, Enoch Aboh, Jan Hulstijn, Judith Rispens, Paola Dussias, and Wander Lowie.

Next, I wish to express my gratitude to the members of the CASLA research group (Cognitive Approaches to Second Language Acquisition), whose member I was during my PhD time. Dear CASLA-ers (Sible, Folkert, Jan, Ineke, Lissan, Silvia, Elisabetta, Tessa, Camille, Rose, Sybren and others), thank you for the many inspiring feedback and discussion sessions we have had. Thank you for the positive and enthusiastic atmosphere that made my PhD much more enjoyable.

My PhD time would have probably been a sad one without my two great office mates, Brechje and Hernán, and our trilingual office, where we did not have any clear pattern as to which language we used (at least not that I have found it, but of course, more research on that is needed). Brechje, my dear office mate and parnymph, thank you for your kindness and friendship, as well as for being a big support during my PhD and at the time when we together thought a course for the first time. Hernán, my dear office mate and fellow animal and plant lover, your view of the world, passion for nature, and your vegan heart brought a lot of brightness into our office, and also many reasons to stop working and start chatting. Thank you for your generous statistical advice, and also for your ingenious idea to convert our office window into a pigeon feeding place. I only wish it had occurred to you earlier than in the last few months of my PhD!

I wish to thank several other PhD colleagues who contributed to my happiness, motivation, and thereby successful completion of this PhD project: Patrick, Camille, Jelke, Matthias, Margreet, Natalia, Caitlin, Tessa, Rosa, Sybren, Sanne, Alaitz, Seid, Tiffany, Jeroen, Bibi, Jasmin, Merel, and Jing. I feel especially grateful to Camille, Patrick, Jelke, and Matthias who were a great support for me during my PhD time. Camille, my dear friend and parnymph, we have had many deep, psychological talks that I enjoyed so much and that I found support in. Patrick, thanks for bringing so much fun and laughter into my PhD time, and thanks for being one of the first people (together with Camille and Jelke) who spoke Dutch and only Dutch to me, and even more importantly, who dared to listen to my Dutch. Jelke, thanks for your generous help, support, and kindness, and also for encouraging my Dutch learning.
Next, what would my PhD be if it was not for the great time I spent living in the Waddendijk house? I shared the house with many housemates over the years and I met many great people, but a few of them left such a deep trace to be mentioned in this short section of my dissertation: Linli, Danis, Aris, Leonardo, Coen, Giovanni, Shengyang, and Lorena. Guys, thank you so much for your friendship, support, fun, laughter, dinners, and parties. You made my PhD time so much brighter and therefore you deserve many many thanks. Linli, thank you for your wonderful friendship, endless optimism, and smiles. Leonardo, your positive energy and love for life and people were contagious, as well as your laughter and crazy dancing! Aris, thanks for the following things you brought into our house: positive energy, crazy masks at dinner parties, and an important date: 4 May, birthday of Star Wars, when Star Wars cupcakes must be made!

I also wish to thank my friends spread all over Amsterdam, Groningen, the Netherlands, Serbia and the world, who even if not in Amsterdam, kept close touch with me and cheered up my PhD time: Ana, Jelena, Srdan, Aida, Dordo, Marava, Lela, Neda, Dunja, Marijana, Vesna and others. Thank you for your generous friendship, help, and support! Also big thanks to Tanja, Nataša, Sabina, and Maja, my wonderful Serbian professors who inspired me, introduced me to the world of linguistic research and wholeheartedly encouraged me to pursue a research career. Also, the research Master program at the University of Groningen, and several professors, Petra, Angeliek, Wander, Kees, and Jacqueline deserve a special place for making me so enthusiastic about research in linguistics.

Finally and most importantly, I want to mention my family and especially four people without whom this final section of my PhD would definitely not exist, people who believed in me and who closely supported me at every single step I took. Mom, thank you for your endless love, support, advice and help. I wish you had been closer to me all these years, but thanks to your love, even through Skype, I felt like at home. Dad, thank you not only for your support and love, but also for making me look at my PhD and school in general from a funnier and less serious perspective. My dear sister and best friend, Sanja, there are no words I can thank you for your support, your readiness to do anything for me, and the most sincere friendship one can have. And last, thank you Francesc, thank you for being close to me all these years, for your love, support, and care. Especially thank you for valuing and encouraging the unserious, funny side of me, and for making me forget my PhD by taking me to mountains, deserts, and islands.
Acknowledgements
# List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACH</td>
<td>aptitude complex hypothesis</td>
</tr>
<tr>
<td>AgrFLK</td>
<td>knowledge of foreign languages with agreement</td>
</tr>
<tr>
<td>ATIs</td>
<td>aptitude by treatment interactions</td>
</tr>
<tr>
<td>BDS</td>
<td>backward digit span</td>
</tr>
<tr>
<td>CI</td>
<td>confidence interval</td>
</tr>
<tr>
<td>DOM</td>
<td>differential object marking</td>
</tr>
<tr>
<td>DS</td>
<td>digit span</td>
</tr>
<tr>
<td>ERPs</td>
<td>event-related potentials</td>
</tr>
<tr>
<td>FDS</td>
<td>forward digit span</td>
</tr>
<tr>
<td>FLK</td>
<td>knowledge of all foreign languages</td>
</tr>
<tr>
<td>GJT</td>
<td>grammaticality judgment task</td>
</tr>
<tr>
<td>GLM</td>
<td>generalized linear model</td>
</tr>
<tr>
<td>GLMM</td>
<td>generalized linear mixed model</td>
</tr>
<tr>
<td>L1</td>
<td>first language</td>
</tr>
<tr>
<td>L2</td>
<td>second language</td>
</tr>
<tr>
<td>M</td>
<td>mean</td>
</tr>
<tr>
<td>NWR</td>
<td>non-word repetition task</td>
</tr>
<tr>
<td>OR</td>
<td>odds ratio</td>
</tr>
<tr>
<td>PI</td>
<td>processing instruction</td>
</tr>
<tr>
<td>RT</td>
<td>reaction time</td>
</tr>
<tr>
<td>SD</td>
<td>standard deviation</td>
</tr>
<tr>
<td>SLA</td>
<td>second language acquisition</td>
</tr>
<tr>
<td>SRT</td>
<td>serial reaction time task</td>
</tr>
<tr>
<td>SSH</td>
<td>shallow structure hypothesis</td>
</tr>
</tbody>
</table>
List of abbreviations
Author contributions

Chapter 1
Written by Maja Ćurčić and revised based on the feedback of Sible Andringa and Folkert Kuiken.

Chapters 2, 3, 4, and 5
The research questions of this dissertation were posed by Maja Ćurčić, under the supervision of Sible Andringa and Folkert Kuiken. These questions were partially based on a project proposal written by Sible Andringa. Maja Ćurčić designed the experiments, made language exposure and test materials, recruited and tested participants, statistically analyzed data, and wrote reports of the studies. Sible Andringa and Folkert Kuiken acted as supervisors of the project and provided feedback on the experimental design of the studies, materials used, data analysis procedures, and draft versions of the articles. Studies presented in Chapters 2, 4, and 5 have been submitted for publication, with Maja Ćurčić as the first author.

Chapter 6
Written by Maja Ćurčić and revised based on the feedback of Sible Andringa and Folkert Kuiken.
Author contributions
Chapter 1: Introduction

Adults show substantial individual differences when trying to learn a new language. They differ in the speed at which they learn, but also in the final success they achieve (Sanz, 2005). These differences are observed by language teachers and laymen alike, and they appear both when learners receive classroom instruction and when they learn language through naturalistic exposure outside the classroom. In the past few decades, language researchers became increasingly interested in what underlies these individual differences. This increased interest led to the enrichment of theory and experimental research that together try to explain where differences in language learning and attainment come from. Identifying factors that can explain such differences is both theoretically and practically relevant. It can help us understand what mechanisms underlie the process of adult language learning, and this knowledge can then be used to improve language instruction.

Thanks to extensive research, it has become clear that many different factors are involved in language learning, and that they interact in very complex ways. These factors can be broadly divided into external and internal ones (Sanz, 2005). External factors could be seen as pertaining to the environment and the language itself. These are for example the characteristics of the language input that learners hear, or the instruction that they receive during the learning process. These general factors can be broken down into different subcomponents that can be studied on their own. For instance, when studying the role of input in language learning, researchers look at how different characteristics of linguistic input, such as complexity, salience, frequency, or regularity, may speed up or slow down learning. Internal factors involved in language learning pertain to learners themselves. These are motivation, personality, cognitive abilities or aptitudes, learning styles and preferences. For instance, when studying how learners’ cognitive abilities influence learning success, researchers look at how learmers’ performance on tests of working memory, analytical ability, statistical learning ability, etc. may or may not predict their performance on language tests. Internal and external factors interact in complex ways, and studying these interactions can be invaluable for better understanding second language (L2) learning processes (DeKeyser, 2012; Sanz, 2005). Research that investigates these interactions, for instance, looks at how certain aptitudes (e.g., working memory, analytical ability) may be involved to a different extent in learning from differential input or instruction (e.g., explicit vs. implicit instruction) – aptitude by treatment interactions (ATIs), or it looks at whether differential aptitudes may be responsible for
learning at different ages (e.g., children vs. adults) – aptitude by age interactions (DeKeyser, 2012).

Broadly speaking, this dissertation focuses on two factors that are involved in explaining differences in adult learning of grammatical patterns: the characteristics of the input and learners’ aptitudes. We investigate these factors by exposing learners to a novel language and by looking at how input characteristics, learners’ aptitudes, and the potential interactions between them explain learning success. We also look at how the involvement of input characteristics may differ for learners who develop awareness of the pattern during language exposure and those who do not. To test learners’ language performance, we use grammaticality judgment tasks (GJTs), as well as an unobtrusive measure of learners’ language processing: the visual world eye-tracking paradigm. The target structure we employ in this dissertation is an agreement pattern between determiners and nouns.

In the remainder of this introductory chapter, we will discuss the most important topics of the dissertation, and we will explain how previous theoretical and experimental research motivated our research questions.

1.1 Input characteristics and language acquisition

Linguistic input is essential for language learning, and no linguistic theory denies it (Sanz, 2005). However, linguistic theories differ with respect to their views on the exact role of input in language learning. For instance, according to nativist UG approaches (Schwartz, 1993; White, 1989), input serves as a trigger for correct parameter setting, and language learning is largely internally driven. On the other hand, usage-based and constructivist approaches give input a more central role: language knowledge develops entirely from the input, which is why characteristics of the input may affect learning (Bybee, 2008; Goldberg, 2006; N. C. Ellis, 2002; Tomasello, 2003).

In developmental psychology, Saffran and colleagues started a whole new field focusing on the statistical properties of the input. For instance, early studies in this field provided evidence that both young children and adults use subtle distributional information in the input to discover word boundaries in running speech (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). This line of experimental research is characterized by the use of artificial languages that do not carry any meaning, carefully controlled language exposure, and subtle manipulations of input distributions to see how these affect learning (for reviews, see Gómez, 2007; Gómez & Gerken, 2000).

In the field of first and second language acquisition, researchers became increasingly interested in which input properties influence language learning and in what ways. Usage-based approaches proposed that the properties of the input, such as complexity, frequency, salience, skewedness, reliability, etc., can
influence both first (L1) and second (L2) language learning (Bybee, 2008; N. C. Ellis, 2002). Given that input properties can influence language learning outcomes, and given that no two learners receive identical input during their learning process, input is probably an important source of differences between L2 learners. Research on how characteristics of language input affect L2 learning is also practically relevant since it can help in shaping and adapting teaching materials and classroom language input in order to maximize learning. Therefore, researchers started conducting carefully designed experimental studies, in which they manipulated characteristics of linguistic input (e.g., salience, complexity, skewedness, frequency) in order to see how these affect learning. Such research has been done on both L1 and L2 language acquisition, and very often learning of grammatical patterns has been the focus. For instance, several studies within the constructivist approach looked at whether patterns in the L1 and the L2 are more easily learned from input that features a very frequent, prototypical exemplar of a pattern (i.e., skewed input) or from input where all exemplars have equal frequency (i.e., balanced input) (e.g., Casenhiser & Goldberg, 2005; McDonough & Nekrasova-Becker, 2014; Year & Gordon, 2009). However, the number of studies that manipulated language input to investigate how such manipulations affect learning is still relatively small. More research in this direction is needed, not only for theoretical reasons, but also given its practical relevance and potential to generate knowledge on how input should be shaped in order to maximize learning and improve language instruction.

This dissertation focuses on an input feature that has received little attention in experimental research so far: input reliability or consistency. In L1 research, there is evidence that unreliable or irregular input can delay acquisition of certain structures (e.g., Costa, Fiéis, & Lobo, 2015; Miller, 2007; Miller & Schmitt, 2012). Miller and Schmitt (2012), for example, found that children who were exposed to Chilean Spanish were delayed in their comprehension of plural morphology compared to Mexican Spanish children. They hypothesized that this was the case because in Chilean Spanish plural marker /s/ is sometimes omitted in pronunciation, which makes the input unreliable compared to Mexican Spanish, where the plural marker is always pronounced. In the field of second language acquisition (SLA), there has also been some interest in reliability as an important characteristic of input that may potentially affect L2 learning outcomes. Several researchers have acknowledged the distinction between reliable or categorical and unreliable or probabilistic rules, and they recognized that these features may lead to differences in learning (e.g., DeKeyser, 1994; Hulstijn & De Graaff, 1994). Also, several experimental studies included input reliability as a variable in their experimental designs. For instance, DeKeyser (1995) looked at whether reliable and unreliable rules are more easily learned via explicit or implicit instruction.
Hudson Kam and Newport (2005) looked at how adults and children learn unreliable language input, and they found evidence that children regularized unreliable input, while adults reproduced it the way they had heard it. However, none of these studies directly compared learning of patterns from reliable and unreliable input, and to our knowledge, there has been no experimental research into how input reliability affects learning of L2 patterns. This dissertation aims to fill in that gap. Additionally, we are interested in whether learners show individual differences in their ability to deal with input unreliability.

1.2 Cognitive aptitudes and second language acquisition

Cognitive abilities or aptitudes are one of the most frequently investigated personal characteristics in individual differences research, probably because knowing which abilities are involved in language learning may have substantial theoretical and practical relevance. Research on aptitudes has the potential to inform the SLA theory on the processes that are relevant for language learning, which in turn may inform teaching practices on the most effective ways to teach languages and adapt language instruction to learners. Although there has been much research on this topic, it is still not entirely clear which cognitive aptitudes are involved in learning of particular linguistic structures, in what way, and under which conditions.

In the early days of aptitude research, Carrol and Sapon (1959) defined language learning aptitude as “basic abilities that are essential to facilitate foreign language learning” (p. 14). In this period, the goal of aptitude-related research was to quickly predict learners' success in acquiring a new language, and this goal led to the construction of different test batteries designed to measure language learning aptitude (e.g., Modern Language Aptitude Test (MLAT), Carrol & Sapon, 1959; Pimsleur Language Aptitude Battery (PLAB), Pimsleur, 1966; The Army Language Aptitude Test (ALAT), Horne, 1971; The Defense Language Aptitude Battery (DLAB), Petersen & Al-Haik, 1976; the VORD, Parry & Child, 1990; The Cognitive Ability for Novelty in Acquisition of Language-Foreign test (CANAL-FT), Grigorenko, Sternberg, & Ehrman, 2000; LLAMA, Meara, 2005). Many of these batteries (especially MLAT and LLAMA) tended to produce fairly high correlations with L2 learning outcomes in classrooms.

In recent years, with the advancement of research on the relation between learning and cognitive abilities, many researchers have agreed that we should talk about language learning aptitudes in plural, rather than treat aptitude as a single, monolithic concept (DeKeyser & Koeth, 2011; Dörnyei, 2006). This is why researchers increasingly started using specific subtests of aptitude batteries or other cognitive abilities tests to investigate how particular
aptitudes predict differences in learning of certain aspects of language. For instance, Doughty et al. (2010) created the High-Level Language Aptitude Battery (Hi-LAB) that measures a number of distinct aptitudes that are potentially involved in language learning at advanced levels, such as rote memory, explicit induction, implicit induction, perceptual acuity, processing speed, inhibition, task-switching, etc. Also, Robinson and Skehan have made some important theoretical advances by recognizing the very complex, multidimensional nature of aptitude in their theories. In his Aptitude Complex Hypothesis (ACH), Robinson (2002, 2005a) draws on Snow (1987, 1994) and suggests the existence of aptitude complexes – combinations of aptitudes that jointly influence language learning under particular learning conditions. Each complex consists of higher, second order cognitive abilities, which in turn consist of primary abilities that jointly facilitate particular aspects of language processing and learning. For example, according to Robinson (2002), explicit rule learning may depend on two higher order abilities: metalinguistic rule rehearsal (relying on grammatical sensitivity and rote memory abilities) and memory for contingent text (relying on working memory for text and speed of working memory for text). Importantly, Robinson’s hypothesis acknowledged that instructional context, and motivational and affective variables influence how aptitudes are used during learning. DeKeyser (2012) further stressed the theoretical and practical relevance of experimentally studying the interactions between aptitudes and other variables such as instructional treatments, age, motivation, etc. Skehan (2002) proposed a model of aptitude where he related particular aptitude components to four main stages involved in L2 grammar learning: noticing (i.e., paying attention to the novel aspect of language), patterning (i.e., generalizing), controlling (i.e., getting to use knowledge in controlled ways), and lexicalizing (i.e., getting to use knowledge fluently and automatically). For instance, he suggested that noticing may rely on aptitudes such as attention management, working memory, and phonemic coding ability, while patterning may rely on grammatical sensitivity and inductive language learning ability. When we consider all these theoretical insights, a complex picture of aptitude emerges, where the involvement of particular aptitudes depends on instructional treatments, linguistic structures, and stages of learning, which may call for more experimental research that does justice to these complex interactions.

Research so far has mainly focused on studying simple relations between aptitudes and learning certain aspects of language, most often without considering the complex interactions between aptitudes and other variables, such as instruction, motivation, learning stages, etc. Given that this dissertation focuses on the acquisition of a grammatical pattern, we will briefly review the literature on the role of aptitudes in pattern learning. Research findings suggest that several aptitudes may be involved in L2 pattern learning: working
memory, measured by a variety of different tasks such as operation span task, reading span task, letter-number ordering task, and others (e.g., Denhovska, Serratrice, & Payne, 2015; Robinson, 2005b; Sagarra & Herschensohn, 2010; Tagarelli, Borges-Mota, & Rebuschat, 2011); analytical ability or the ability to induce grammar rules from the input, often measured by subtests of MLAT and PLAB (e.g., Erlam, 2005; Harley & Hart, 1997; Robinson, 2005b); and statistical learning ability or the ability to become sensitive to underlying patterns in the input, measured by visual serial reaction time tasks (SRT) or tests tapping into learners' ability to detect verbal nonadjacent dependencies (e.g., Brooks & Kempe, 2013; Granena, 2013; McDonough & Trofimovich, 2016). The studies listed above have shown that learners who score higher on measures of working memory, analytical ability, or statistical learning ability also demonstrate better performance on language tests, which may imply that better cognitive aptitudes lead to better language learning.

Knowing which aptitudes are involved in learning a particular aspect of language gives the possibility of adapting language instruction in such a way to facilitate the use of the relevant learning mechanisms. For instance, in their working memory model, Baddeley and Hitch (1974) suggest that rehearsal is a component of working memory, and the activity of rehearsing input helps in keeping input longer in short-term memory. This implies that by encouraging learners to rehearse language input, it should be possible to help them keep input longer in their working memory, and learn better as a result. Indeed, there is evidence that instruction with rehearsal leads to better vocabulary learning (Ellis & Beaton, 1993; Papagno, Valentine, & Baddeley, 1991; Seibert, 1927; Yoshida & Fukada, 2014), but the evidence is still scarce for learning of grammatical patterns. Research on how rehearsal may influence pattern learning can provide further evidence for the involvement of working memory in pattern learning. In addition, it has the potential to identify the ways in which instruction can be adapted so as to make a better use of the learning mechanisms that are relevant for learning particular grammatical patterns.

In this dissertation, we investigate the involvement of several aptitudes (i.e., working memory, analytical ability, statistical learning ability) in learning a novel L2 pattern. The designs of our studies also allow us to look at aptitude by treatment interactions (ATIs), i.e. whether the involvement of aptitudes in pattern learning differs in different instructional treatments. In addition, we look at whether pattern learning can be improved by helping learners to keep input longer in their working memory through rehearsal.

1.3 Individual differences in awareness during exposure

When learning a new language, adult learners bring their past language experiences into the learning process, which can result in heightened noticing
and awareness of certain L2 features. The role of noticing and awareness in L2 learning has been one of the more intensively discussed topics in second language acquisition (SLA) research, forming the basis of research into explicit and implicit learning. The interest in the topic of awareness in SLA dates back to Krashen (1977, 1979, 1981), who argued that L2 learning relies on incidental process of acquisition, with little role for awareness and explicit learning processes. Later, Schmidt's Noticing Hypothesis (see Schmidt, 1995) and Robinson’s model of the relationship between attention and memory (Robinson, 1995) recognized the central role of noticing and awareness in second language learning.

In the SLA literature, there has been substantial attention for two main methodological problems related to awareness. The first problem relates to how learners’ awareness should be measured (for a discussion see Rebuschat, Hamrick, Riestenberg, Sachs, & Ziegler, 2015), which is crucial for teasing apart explicit from implicit language learning (for discussions on measuring explicit and implicit knowledge, see Andringa & Rebuschat, 2015; R. Ellis, 2005a). Studies have used different methods to measure learners’ awareness, ranging from think-aloud protocols and offline retrospective verbal reports to trial-by-trial source attributions, where learners are asked to report the basis or source of their grammaticality judgments. However, each measure comes with some disadvantages. While think-aloud protocols and source attributions may encourage learners to look for patterns and trigger their awareness, offline verbal reports may lead to underestimating the number of aware learners (Rebuschat et al., 2015). The second problem (noted in Doughty, 2003; Norris & Ortega, 2000) relates to fact that most tasks in SLA research are biased towards the use of explicit knowledge (e.g., GJs). Additionally, such tasks may trigger learners’ awareness.

Learners show substantial individual differences in whether or not they develop awareness of particular input features during language exposure (e.g., Rebuschat et al., 2015; Williams, 2005) However, when interpreting learning outcomes, many studies do not pay enough attention to the role of awareness. Some of these studies make no mention of potential differences in awareness between learners, thereby implicitly assuming that there are no differences (e.g., Goldberg, Casenhiser, & Sethuraman, 2004; McDonough & Nekrasova-Becker, 2014; Year & Gordon, 2009), whereas other studies report the existence of differences, but fail to take them into account when presenting and interpreting results (e.g., Safran, Newport, Aslin, Tunick, & Barrueco, 1997). Differences in the occurrence of awareness are only systematically reported in studies that aim to investigate implicit learning or awareness, and because of this need to classify learners as aware or unaware (e.g., Rebuschat et al., 2015; Williams, 2005). Because awareness heavily affects learners’ performance and behavior on tests, studies investigating language learning after exposure cannot
be fully interpreted if awareness-related differences between learners are not taken into consideration.

In this dissertation, participants were exposed to a novel miniature language, and GJT's were used to measure their pattern knowledge. Given that the exposure paradigm we used may lead to substantial individual differences in the occurrence of awareness, and given that GJT's may also trigger search for patterns and awareness, we systematically looked at awareness-related differences between participants. We also used these differences to better interpret the results throughout this dissertation.

### 1.4 Individual differences in L2 processing

Language learners have also been shown to exhibit substantial individual differences in the extent to which they are able to efficiently use their language knowledge in time during comprehension (Roberts, 2012). Measures of language processing (i.e., eye-tracking, event-related potentials - ERPs, self-paced reading/listening) tap into learners' ability to use language quickly as it unfolds in time. The advantage of some of these measures is that they are unobtrusive. For instance, when using the eye-tracking measure, learners can be exposed to natural language, and learners' knowledge of certain structures can be assessed without them knowing what aspect of language is being tested; this is difficult to achieve with many other language measures, such as GJT's, where ungrammatical items are used, which may make learners aware of the target structures.

Online processing measures have been extensively used in L2 research to investigate how learners process particular L2 structures compared to native speakers. It has been found that L2 learners show similar processing of lexical and semantic information as native speakers (e.g., Roberts & Felser, 2011; Williams, 2006), but that differences and difficulties mainly lie in the domain of grammar (Havik, Roberts, Van Hout, Schreuder, & Haverkort, 2009; Roberts, 2012). Language proficiency and similarity between L1 and L2 structures have been identified as important factors that at least partially explain why some learners get closer to nativelike processing than others (e.g., Dussias, Contemori, & Román, 2014; Dussias, Valdés Kroff, Guzzardo Tamargo, & Gerfen, 2013; Foucart & French-Mestre, 2011; Sabourin & Stowe, 2008). There is little research on the role of other factors in explaining individual differences in online processing, such as aptitudes, awareness, and kind of instruction. More recently, studies using online processing measures have increasingly started to look at what kind of processing learners can show after brief exposure to novel linguistic structures (e.g., Andringa & Curcic, 2015; Batterink & Neville, 2013; Davidson & Indefrey, 2009; Marsden, Williams, & Liu, 2013; Morgan-Short, Sanz, Steinhauer, & Ullman, 2010). For instance, Davidson and Indefrey (2009)
showed that learners can come to process adjective declension in their L2 in ways similar to native speakers after a short explicit instruction. This implies that online processing measures have a potential for studying immediate outcomes and effectiveness of L2 instruction. This is especially the case when it comes to relatively unobtrusive measures, such as eye-tracking, which allow assessing learners’ linguistic processing without drawing their attention to the linguistic aspects under investigation.

In this dissertation, in addition to GJTs, we used the visual world eye-tracking paradigm to assess participants’ processing of the target structure. We aimed to see whether learners can develop nativelike processing of determiners after a short exposure to a novel language, and how their aptitudes, awareness, and kind of instruction may modulate their processing abilities.

**1.5 Goals and research questions**

Broadly speaking, this dissertation focuses on differences in learning of determiner-noun agreement pattern in adult second language acquisition, and on some factors that may explain these differences, such as input characteristics and aptitudes. The goal of the dissertation is to answer the following main research questions:

1) *Does input reliability affect L2 pattern learning, and in what way?* (Chapters 2 and 3)

2) *Can different cognitive aptitudes (i.e., working memory, analytical ability, and statistical learning ability) predict L2 pattern learning? Does aptitude involvement differ for different instructional treatments?* (Chapters 2, 3, and 4)

3) *Can input rehearsal improve L2 pattern learning by helping learners to keep input longer in their working memory?* (Chapter 4)

4) *Can L2 learners develop online processing of determiners after a short exposure to the target structure? Do factors such as cognitive aptitudes, awareness, and type of instruction explain differences in this processing?* (Chapter 5)

Research question 1 is related to the role of input characteristics in L2 pattern learning, and explaining differences between learners. More specifically, the question focuses on the role of unreliable input with exceptions on learning determiner-noun agreement pattern. L1 research has already provided some evidence that unreliable input can disturb pattern learning in
children, but to our knowledge, no experimental studies have investigated this possibility in relation to L2 learning.

Research question 2 concerns the role of cognitive aptitudes in explaining and predicting individual differences in L2 pattern learning. We focus on several aptitudes that previous research has addressed: working memory, analytical ability, and statistical learning ability. This dissertation aims to provide further evidence for the involvement of these aptitudes in learning the determiner-noun agreement pattern, and it also looks at whether there are interactions between aptitudes and instructional treatments (i.e., reliable vs. unreliable input instruction; rehearsal vs. non-rehearsal instruction).

Research question 3 addresses the possibility of improving language learning through instruction activities that rely on particular aptitudes. More specifically, we investigate whether pattern learning can be improved through instruction that contains rehearsal, which is known to help keeping input longer in working memory.

Finally, research question 4 investigates whether learners can come to process L2 structures in similar ways as native speakers after a very short exposure. Here we also look at several factors that may play an important role in explaining individual differences in L2 processing but have not received enough attention in research so far. More specifically, we look at the relation between learners’ processing of determiners and their cognitive aptitudes (i.e., working memory and analytical ability), awareness at different levels, and whether or not instruction contains input rehearsal.

1.6 Outline of the dissertation

This dissertation includes three experimental studies reported in four chapters. Chapter 2 reports results of a study that investigates the role of unreliable input and several aptitudes in L2 pattern learning. Chapter 3 reports findings of an experimental study that uses longer exposure to replicate the Chapter 2 findings and provide stronger evidence of the effects observed. In addition, Chapter 3 explores how implicit learning and awareness may develop over time, and how these two types of knowledge may influence each other. Chapter 4 reports the results of a study that investigates if rehearsal can lead to improved pattern learning, as it helps learners to keep input longer in their working memory. Chapter 5 reports the eye-tracking data collected in the previously reported studies. In this chapter, we look at whether learners can develop nativelike processing of determiners after a very short exposure, and how learners’ aptitudes, awareness, and type of instruction account for individual differences in this processing. Finally, in Chapter 6, we bring together and discuss the results of the dissertation, and we point out some
methodological considerations, directions for future research, and practical implications of the dissertation.

There is substantial overlap between the method sections in Chapters 2, 3, 4, and 5 because the studies presented in this dissertation share similar methodology and have been written in such a way that they can be read independently from each other.
Chapter 1
Chapter 2: The role of unreliable input in L2 pattern learning

Abstract

This study investigates the effects of input reliability on acquiring a novel pattern in adult second language acquisition. It also explores whether learners’ aptitudes, such as working memory, analytical ability, and statistical learning ability can explain differences in pattern learning, and whether this may differ for learning from reliable and unreliable input. Participants in the study received auditory exposure to a miniature language based on Fijian and determiner-noun agreement pattern, after which we measured their target pattern knowledge using two grammaticality judgment tasks, one with familiar and one with novel nouns. Results showed that some learners became aware of the target pattern, whereas the others did not, and there were substantial differences in pattern acquisition. Also, implicit learning without awareness has been found. Unreliable input disturbed pattern learning, but only when learners were aware of the target structure. We found links between pattern learning and both working memory and analytical ability. There was also some evidence that involvement of working memory in pattern learning depended on input reliability, i.e. working memory predicted pattern learning from reliable input better than from unreliable input.

2.1 Introduction

When exposed to a language, learners gradually abstract the underlying patterns from the input, and they differ substantially in their ability to do so (Rebuschat & Williams, 2012). Many external and internal factors, as well as complex interactions between them are responsible for these differences (Sanz, 2005).

One of the important external factors that affect pattern abstraction is input and its characteristics (Bybee, 2008; N. C. Ellis, 2002). Investigating which input properties are more or less conducive to pattern learning is not only theoretically, but also practically relevant since it could have important implications for second language teaching and tailoring language instruction and input to improve learning outcomes. This is why studies that experimentally investigate how input properties affect learning outcomes are

*A slightly modified version of this chapter has been submitted as Curcic, M., Andringa, S., & Kuiken, F. (under review). Exceptions disturb learning of linguistic patterns in adult second language acquisition: An individual differences approach.
especially important. Such studies exist, but they are still relatively scarce, and they have addressed only some of the many potentially relevant input features. The main goal of this study is to investigate how adult L2 learners abstract a linguistic pattern from unreliable and reliable input (i.e., input with and without exceptions). Thereby, we aim to shed light on how input reliability affects pattern learning, a topic that has been investigated in first language (L1) learning (e.g., Costa et al., 2015; Miller & Schmitt, 2012), but hardly in second language (L2) research.

Internal factors, which pertain to learners themselves, also influence pattern abstraction and language learning outcomes. One of the most important and most intensely studied internal factors are learners’ cognitive abilities or aptitudes (Dörnyei, 2006). It has also been recognized that the involvement of aptitudes in language learning may depend on other factors, such as linguistic structures being learned, instructional treatments, learning stages, etc. (DeKeyser, 2012; Robinson, 2002, 2005a; Skehan, 2002). However, while research on these complex interactions has large theoretical and practical relevance in understanding and promoting L2 learning, there are still relatively few studies in this direction. The present study aims to contribute to the existing body of aptitude research by trying to relate learners’ performance on several aptitude tests to their pattern learning. Additionally, the design of the study allows us to look at whether aptitude involvement may differ for learning from reliable vs. unreliable input (i.e., aptitude by treatment interactions – ATIs).

2.1.1 Input characteristics and L2 pattern learning

All linguistic theories agree that input plays a crucial role in language learning (Sanz, 2005). However, the point of disagreement comes from different views on the exact role that input plays in the learning process. While UG approaches see input as a trigger of internally built parameter settings (Schwartz, 1993; White, 1989), usage-based and constructivist approaches regard it as the basis of all language learning (Bybee, 2008; Goldberg, 2006; N. C. Ellis, 2002; Tomasello, 2003).

In the field of developmental psychology, Saffran and colleagues investigated how children and adults learn word segmentation from running speech thanks to distributional properties of the input (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). This was the start of much research into statistical learning, where researchers took interest in how different input distributions affect language learning outcomes, which is why they employed miniature artificial languages and manipulated their features to assess their influence on language learning (for reviews see Gómez, 2007; Gómez & Gerken, 2000).
Given the crucial role of input in L2 learning, it is surprising that there have been relatively few experimental studies in the second language acquisition (SLA) field that manipulated input characteristics in order to assess how such manipulations affect L2 learning. From a theoretical point of view, usage-based approaches to language acquisition have made some important advances in hypothesizing that different input characteristics (e.g., complexity, frequency, reliability/consistency, salience, skewedness) may affect the process of pattern abstraction from input (Bybee, 2008; N. C. Ellis, 2002). These insights have triggered experimental studies that manipulated linguistic input in subtle ways in order to shed light on how particular input characteristics affect pattern learning. It is surprising that the number of such studies is scarce given their potential practical relevance for improving language instruction. Most of the existing studies have been conducted on the effectiveness of skewed vs. balanced input in abstracting novel patterns in L1 and L2 (e.g., Casenhiser & Goldberg, 2005; Goldberg et al., 2004; McDonough & Nekrasova-Becker, 2014; Year & Gordon, 2009). Skewed input includes a highly frequent exemplar of a particular structure with a prototypical meaning, whereas in balanced input, all exemplars are equally frequent. In L2, there is some evidence that skewed input facilitates pattern learning compared to balanced input (e.g., Goldberg et al., 2004), but findings on this topic have been mixed given that some studies found the opposite result, i.e. that balanced input was more effective than skewed input (e.g., McDonough & Nekrasova-Becker, 2014; Year & Gordon, 2009).

There is also some research on how complexity of linguistic structures affects learning and effectiveness of explicit and implicit instruction (e.g., De Graaff, 1997; Housen, Pierrard, & Vandaele, 2005; Spada & Tomita, 2010). Studies on complexity have compared learning of complex vs. simple structures, but serious challenges in this line of research are the existence of many different definitions and measures of complexity (Housen & Kuiken, 2009; Spada & Tomita, 2010), as well as difficulties to isolate effects of complexity from other features in which complex and simple rules often differ, such as salience and frequency.

The effects of input reliability/consistency on acquisition of linguistic structures have been studied to some extent in first language acquisition (e.g., Costa et al., 2015; Johnson, 2005; Miller, 2007; Miller & Schmitt, 2012), but very little in the L2 acquisition field. The L1 studies that addressed the topic of input reliability were conducted in naturalistic settings and did not manipulate language input in controlled ways. Most of these studies suggested that unreliable input delays acquisition of language structures, such as plural marking and clitic placement (e.g., Costa et al., 2015; Miller, 2007; Miller & Schmitt, 2012). For instance, Miller and Schmitt (2012) compared plural marking acquisition of children exposed to Chilean Spanish, in which plural marking is unreliable in the input, with plural marking acquisition of children...
exposed to Mexican Spanish, in which plural marking is reliable and consistent. They found that unreliable plural marking in input delayed children’s ability to use plural morphology in comprehension.

SLA researchers have observed that linguistic patterns can differ in terms of their reliability, i.e. whether they are reliable/categoric or unreliable/probabilistic (e.g., DeKeyser, 1994; Hulstijn & De Graaff, 1994), and they have also recognized that differences in reliability of rules may lead to differences in their learning (e.g., Hulstijn & De Graaff, 1994). There has been some experimental research on the effectiveness of explicit and implicit instruction for the learning of categorical vs. unreliable rules in L2 (DeKeyser, 1995), and also some research on whether adult and child learners regularize or reproduce unreliable input (Hudson Kam & Newport, 2005). However, to our knowledge, there has been no experimental research that systematically investigated how input reliability affects L2 pattern learning, i.e. whether exceptions disturb L2 pattern learning or not. Investigating this is important considering that most linguistic structures feature exceptions. Such research could also be pedagogically beneficial in tailoring classroom input so as to optimize learning of particular linguistic structures.

### 2.1.2 Aptitudes and L2 pattern learning

Apart from input characteristics, learner characteristics play an important role in language learning, and this is especially the case in adult L2 acquisition (Dörnyei, 2006). Learners show substantial individual differences in their motivation, personality, learning styles, and also cognitive abilities or aptitudes (Dörnyei, 2006; R. Ellis, 2005).

The relation between cognitive aptitudes and language learning success has been studied in both naturalistic settings, where participants’ performance on aptitude tests is related to their ultimate L2 attainment (e.g., Granena, 2013; Harley & Hart, 1997; Robinson, 1997) and in experimental settings, where learners are briefly exposed to a new miniature language or a semi-artificial language, after which their learning success is related to their performance on aptitude tests (e.g., De Graaff, 1997; Grey, Williams, & Rebuschat, 2015; McDonough & Trofimovich, 2016). Compared to the studies in naturalistic settings, experimental studies provide much better control of the type and amount of input learners are exposed to, which helps to better isolate the role of aptitudes. However, researchers in this line of research often employed either miniature languages that had little or no resemblance to real languages, or miniature languages that were based on learners’ L1 vocabulary, which may have resulted in learners becoming aware that they were learning a non-existing language.
Although the relation between aptitudes and L2 learning has been extensively investigated, there are still very few clear answers. This may be partially due to researchers using different aptitude measures and the fact that it is difficult to be fully certain about which aptitude measures ought to be used or which aptitude measures may predict the type of learning under investigation. Another reason is that the involvement of aptitudes in language learning is not a matter of simple correlation. Several researchers pointed to the complex, multidimensional role of aptitude in language learning, where aptitude interacts with instructional treatments, learning stages, the linguistic structures learned, learners’ motivation, age, etc. (DeKeyser, 2012; Robinson, 2002, 2005a; Skehan, 2002). For instance, Robinson’s Aptitude Complex Hypothesis (2002, 2005a) suggests that different learning situations rely on different aptitude complexes, i.e. combinations of higher order aptitudes that further consist of primary aptitudes. For example, explicit rule learning may rely on metalinguistic rule rehearsal and memory for contingent text, whereas incidental learning from oral input may rely on memory for contingent speech and deep semantic processing. Skehan (2002) argues that different aptitudes may be responsible for learning of grammatical patterns at different stages: initial noticing of relevant aspects of input may rely on learners’ working memory, phonemic coding ability, and their attention management, whereas the next learning stage that leads to extracting patterns may rely on grammatical sensitivity and inductive language learning ability.

Many studies looked at simple relations between aptitudes and L2 pattern learning. Research so far has found that several main aptitudes are involved in L2 pattern learning: working memory (e.g., Denhovska et al., 2015; Robinson, 2005b; Sagarra & Herschensohn, 2010; Tagarelli et al., 2011), analytical ability – the ability to induce grammar rules and consciously extract patterns from the input (e.g., Erlam, 2005; Harley & Hart, 1997; Robinson, 2005b), as well as aptitudes tapping into more implicit cognitive processes, such as statistical/sequence learning ability – the ability to develop sensitivity to underlying patterns in the input (e.g., Brooks & Kempe, 2013; Granena, 2013; McDonough & Trofimovich, 2016). These studies found that if learners performed better on aptitude tests, they also showed better learning outcomes/knowledge of the target L2 patterns. However, it should be noted that a variety of different tests have been used to measure the same or similar aptitudes, and it is often unclear which underlying aptitudes the tests really measure. For instance, working memory has been measured by a variety of different tests, such as reading span task (e.g., Denhovska et al., 2015; Robinson, 2005b; Sagarra & Herschensohn, 2010, Tagarelli et al., 2011), digit span task (e.g., McDonough & Trofimovich, 2016), non-word repetition task (e.g., Grey et al., 2015), operation span task (e.g., Denhovska et al., 2015; Tagarelli et al., 2011), letter-number ordering task (e.g., Tagarelli et al., 2011), and many
others. This is why caution is needed when interpreting findings and drawing firm conclusions about which aptitudes are involved in learning a particular linguistic structure. This situation may also call for more exploratory research in order to accumulate evidence and establish stronger and more reliable links between particular aptitudes and pattern learning.

As suggested in DeKeyser (2012), more research is needed that looks at the interaction between aptitudes and other variables. Several studies have investigated so-called aptitude by treatment interactions (ATIs) (e.g., Denhovska et al., 2015; Erlam, 2005; Robinson, 2005b; Wesche, 1981). Robinson (2005b) showed that learning from the implicit condition was predicted by working memory, whereas learning from the explicit condition was predicted by grammatical sensitivity. Erlam (2005) found a relation between inductive instruction and analytical ability, whereas structured input treatment was related to both analytical ability and working memory. A recent study by Denhovska et al. (2015) showed that working memory was involved in the acquisition of noun-adjective agreement in Russian when learners received input with high type and high token frequency. However, when input featured low type and low token frequency, working memory no longer predicted learning. As pointed out in DeKeyser (2012), our understanding of such aptitude by treatment interactions is still very limited, and their practical relevance is substantial, which is why more research in this area is needed.

2.1.3 The present study

Participants in this study learned a novel miniature language based on Fijian under controlled input conditions. Importantly, the vocabulary did not resemble either the participants’ L1 or possible L2 vocabularies, and learners believed they were learning a real language. The language exposure was auditory, and the target pattern was determiner-noun agreement. In order to investigate the effects of input reliability on learning, we systematically manipulated the input type: half of the learners received the input without exceptions to the target pattern – the reliable input group, whereas the other half heard the input with some exceptions – the unreliable input group.

We employed several aptitude measures that we hypothesized – based on the previous literature – to be relevant to the learning of the agreement structure. We used the following aptitude tests: the LLAMA D sound recognition task (Meara, 2005), tapping into verbal working memory; a probabilistic serial reaction time (SRT) task (Kaufman et al., 2010), a measure of non-linguistic (i.e., visual) statistical learning ability; and two measures of analytical ability: the LLAMA F grammatical inference task (Meara, 2005) and a non-verbal IQ test (Wechsler, 2008). The goal was to see if these aptitudes could account for individual differences in learning the target pattern. The design of the study
also allowed us to look at ATIs, i.e. whether the involvement of the aptitudes depended on instructional treatment (reliable vs. unreliable input). We tentatively hypothesized that learning from unreliable input may rely more on statistical learning ability and memory, given that learning from this kind of input involves tracking of probabilities and memorizing regular items and exceptions. Conversely, we speculated that learning from fully reliable input may allow for better use of learners’ analytical abilities given that reliable input features no exceptions that would disturb analyzing input and extracting patterns.

The learners’ target structure knowledge was tested using two oral grammaticality judgment tasks (GJTs): one with items that were familiar from the exposure and one with novel items. We assumed that good performance on the GJT with familiar items may be based on memory, i.e. item-based knowledge, but potentially also on pattern knowledge. However, good performance on the GJT with novel items could only result from pattern knowledge. Our goal was to answer the following research questions:

1) Does input reliability affect L2 pattern learning, and in what way?

2) Can working memory, analytical ability, and statistical learning ability predict learning of the pattern?

3) Does aptitude involvement differ for different instructional treatments (i.e., reliable vs. unreliable input)?

We also collected eye-tracking data to investigate individual differences in learners’ online processing of the target structure; these results are presented in Chapter 5.

2.2 Method

2.2.1 Participants

Participants in the study were 50 adult native speakers of Dutch, between 19 and 35 years old ($M_{age} = 24.62, SD = 4$). They were students or highly educated adults, without education in linguistics. All participants reported having good hearing, normal or corrected-to-normal vision, no history of dyslexia, and no color-blindness. Each participant was randomly assigned to one of the two instruction groups: the reliable input group (7 males, 18 females, $M_{age} = 24.48, SD = 3.56$) and the unreliable input group (7 males, 18 females, $M_{age} = 24.76, SD = 4.47$).

All learners reported having knowledge of one or more foreign languages. Since prior knowledge of foreign languages can have positive influence on
learning, we wanted to check if the two groups differed in this respect. In a questionnaire, participants were asked to list the foreign languages they spoke, and to assess the level of their knowledge (A1 – C2 on the CEFR scales (Council of Europe, 2011)). Based on learners’ reports, we computed two measures that reflected participants’ knowledge of foreign languages: 1) a measure that took into account the number and level of all foreign languages (foreign language knowledge – FLK) and 2) a measure that took into account the number and level of foreign languages with determiner-noun agreement, i.e. Spanish, French, etc. (agreement foreign language knowledge – AgrFLK). The measures were created by awarding 0.5 points for each level of each language. The average FLK was 8.15 (SD = 3.49), whereas the average AgrFLK was 1.91 (SD = 2.45). A Mann-Whitney U test showed that reliable (M = 7.66, SD = 3.64) and unreliable input learners (M = 8.64, SD = 3.34) did not differ significantly on the FLK measure, i.e. knowledge of all foreign languages (W = 261.5, p = .32, 95% CI [-3, 1]). Also, there was no significant difference between reliable (M = 1.52, SD = 1.96) and unreliable input learners (M = 2.3, SD = 2.84) on the AgrFLK measure, i.e. the knowledge of languages with determiner-noun agreement (W = 257.5, p = .27, 95% CI [-1, 0.0001]).

The experimental session lasted about 1 h and 45 min, and learners received 15 euros for participation. The experiment was approved by the Ethics Committee of the University of Amsterdam.

2.2.2 The target language and the target structure

Participants were exposed to a miniature language that we created by modifying Fijian – an Austronesian language of the Malayo-Polynesian family, spoken in Fiji. The target structure was determiner-noun agreement.

Lexically, the language we created did not resemble Dutch or any of the foreign languages our learners were familiar with. The lexical items were based on Fijian, and the language consisted of 20 nouns, four verbs, six adjectives, and two determiners (Appendix A). Twenty additional nouns were used for the grammaticality judgment task (GJT) with novel items. Nouns were created by adding -is/-uk endings to Fijian nouns, whereas verbs were created by adding a -t suffix to the existing Fijian verbs. Where possible, we took over both the form and meaning of Fijian words. However, if a form for a certain meaning was too complex or potentially reminded of a word from participants’ L1 or L2s, we chose a form of another Fijian word.

The language had Subject – Verb – Object word order, and the linguistic input learners received featured simple phrases and intransitive sentences in

---

1 Whenever t-test assumptions (i.e., normality of distribution and homogeneity of variance) were violated, we used a non-parametric Mann-Whitney U test instead of a t-test.
The role of unreliable input in L2 pattern learning

The present tense, which was expressed by a simple tense, marked by a -t suffix in the third person singular.

Nouns were preceded by a definite determiner that had two possible forms, depending on the grammatical gender of the noun it preceded. Lep was a masculine determiner and preceded nouns ending in –uk (e.g., *lep dawuk* (“the cat”)). Ris was a feminine determiner and preceded nouns ending in –is (e.g., *ris euilis* (“the bicycle”)). Dutch – the participants’ L1 – also has two definite determiners (*de* and *het*), which express a distinction between common (*de*) and neuter (*het*) gender. However, there is no feminine vs. masculine distinction between them, and also no clear determiner-noun agreement (Booij, 2002).

The goal of the language exposure was to teach participants the determiner-noun agreement structure. For the reliable input learners, the agreement pattern featured no exceptions. However, for the unreliable input group, the pattern was reversed for 4 out of the 20 nouns that participants learned: two -is nouns were preceded by *lep* (e.g., *lep vonuis* (“the turtle”)) and two -uk nouns were preceded by *ris* (e.g., *ris touk* (“the goat”)).

### 2.2.3 Language exposure

Before the language exposure started, participants were told that they would be learning a new language by looking at images and listening to sentences describing the images. They were not made aware of the purpose of the experiment and the target structure, but they were informed that their knowledge of the language would be tested from time to time.

Participants did the experiment in a quiet room, and the materials were presented on a computer screen using the E-prime software. The listening materials were presented through loudspeakers. All experimental instructions were written on the computer screen in Dutch, and were also recorded by a female native speaker of Dutch. The experimental materials were presented auditorily, and were recorded by a female native speaker of Serbian to give the language a foreign feel. The images used in the experiment were retrieved from the Clipart image database (Clipart, Vector Graphics, and Illustrations, 2014) and edited in Photoshop.

**Noun learning and assessment**

The exposure started with a noun learning phase, whose purpose was to make sure that all learners acquired the nouns equally well before receiving further exposure to the language. This was important in order to later create equal opportunities for target structure learning and to avoid test results being influenced by differences in the knowledge of nouns.
In this phase, learners were only exposed to 20 nouns of the new language, and were not exposed to the determiner-noun agreement structure. In every trial, they saw a simple black-and-white image and heard the noun denoting the object in the image. Every noun appeared six times. The order of presentation of the trials was identical for all learners.

The noun learning phase was immediately followed by a test, in which we assessed learners’ knowledge of the nouns. In each trial, learners saw four images on the screen and heard a noun. Their task was to click on the correct image that corresponded to the noun they had heard. Each noun appeared as the target noun in two trials, whereas the distractor nouns were randomly chosen by the program. The test continued until learners had 100% accuracy on the test. This means that all learners completed a minimum of 40 trials that were identical for all of them. If learners made mistakes on some of the nouns, after the 40 obligatory trials, they would receive additional trials with the wrongly identified nouns as targets, and this would continue until they reached 100% accuracy. Learners were given feedback about the accuracy of their responses so that they could continue learning.

Given that the noun test measured how successfully learners could form associations between nouns and their referents, we also used it as a measure of learners’ rote memory ability. However, since the minimum number of items that learners needed to complete was set by the test (i.e., 40), it may not have measured the full range of differences in rote memory abilities present in the sample. The results of the noun test are presented together with other aptitude measures in subsection 2.2.6.

**Target structure exposure**

In this part of the exposure, participants received more complex linguistic input that included the determiner-noun agreement structure (Appendix B). In every trial, the participants saw an image representing a simple object or an intransitive activity, and they heard a phrase or sentence describing the image. In addition to the 20 nouns from the noun learning phase, this part of the exposure featured three different intransitive verbs, copula be, six adjectives, and two determiners. The items fell into one of the following four structural categories:
The role of unreliable input in L2 pattern learning

1) Determiner + noun

![Chair](image1)

*Ris salis*

"The chair"

2) Determiner + adjective + noun

![Chair](image2)

*Ris matene salis*

"The blue chair"

3) Intransitive sentences: determiner + animate noun + verb_{intransitive}

![Dog](image3)

*Ris burogis sisilit.*

"The dog is swimming."

4) Intransitive sentences: determiner + inanimate noun + is + adjective

![Chair](image4)

*Ris salis na duka.*

"The chair is broken."

Learners heard a total of 308 trials, and the exposure was divided into two parts with a break in between so that the learners would not lose concentration. The first part consisted of 176 trials and lasted 15 minutes.
Eighteen out of 20 nouns appeared eight times, and two nouns were overrepresented, i.e. more frequent, so that the input would resemble the natural input more (Ellis & Ferreira-Junior, 2009): one feminine noun (burogis – “dog”) and one masculine noun (dawauk – “cat”). These nouns appeared 16 times each. The second part of the exposure lasted 10 minutes, and learners saw 132 trials. The nouns that were not overrepresented appeared 6 times each, whereas the overrepresented nouns appeared 12 times each. The order of the presentation of the trials was identical for all learners.

Learners received equal exposure to both lep – uk and ris – is agreement patterns. The two determiner categories had the same number of overrepresented vs. underrepresented nouns, animate vs. inanimate nouns, non-biological vs. biological gender nouns (i.e., boy, girl, man, woman).

Four out of 20 nouns – two feminine and two masculine – were exceptions to the determiner-noun agreement pattern, but this was the case for the unreliable input learners only. These were so-called inconsistent items. For instance, the noun touk was presented with the masculine determiner lep to the reliable input learners, whereas it was presented with the feminine determiner ris to the unreliable input learners. The inconsistent items were gradually introduced after 30 trials.

2.2.4 Oral GJTs

After the exposure, we administered two grammaticality judgment tasks (GJTs) to test learners’ knowledge of the target structure. The first GJT featured nouns that were unfamiliar to the learners from the exposure phase (Appendix C), whereas the second one featured nouns that participants had been exposed to (Appendix D). We were aware of the possibility that some learners may search for patterns during the GJT tasks, and that during the GJT with familiar nouns it may be easier to find the pattern. Having this in mind, we administered the GJT with novel nouns before the GJT with familiar nouns.

In every trial, learners saw a simple image and heard a short phrase consisting of a determiner and a noun. Before the GJT with novel nouns, learners were told that they would hear phrases that they had not heard before, whereas before the GJT with familiar nouns they were told that the phrases would be familiar to them. Learners were asked to decide whether the phrase they heard was good or not in the language they had learned. They were allowed to re-play the phrase as many times as they wanted and were encouraged to use their intuition or to guess if they were not sure about the correct answer.

In both GJTs, we systematically varied the noun type – is nouns vs. -uk nouns, and the grammaticality of the phrase – grammatical or ungrammatical. This gave us four experimental conditions, with 10 items per condition.
Ungrammaticality of the phrases always stemmed from incorrect determiner-noun combination. The GJT with familiar nouns featured inconsistent nouns as well, and accuracy on them always reflected what learners had heard in the exposure phase. For instance *lep touk* was correct for reliable input learners, but incorrect for unreliable input learners. Test items were presented in a fixed random order that was identical for all participants.

2.2.5 Debriefing
After the language exposure and tests, we debriefed learners using a protocol in order to find out whether they noticed the target pattern, and if so, at which stage in the experiment (i.e., during exposure, GJT with familiar nouns, GJT with novel nouns, or during debriefing). We also asked learners about what kind of knowledge they relied on during each of the two GJTs. If learners could at least partially verbalize the target agreement pattern, they were classified as aware. If learners did not report any target pattern awareness, we told them that there was a pattern in the language, and invited them to guess what it was. If they still showed no awareness, we explained the target pattern, and we told them that the language was a non-existing language based on Fijian. We used debriefing results to classify learners as aware or unaware of the target pattern in different stages of the experiment.

2.2.6 Aptitude measures
We administered several different aptitude tests: the LLAMA D sound recognition task (Meara, 2005) – a measure of verbal working memory, the LLAMA F grammatical inference task (Meara, 2005) – a verbal measure of analytical ability, the IQ test (Wechsler, 2008) – a non-verbal measure of analytical ability, a serial reaction time task (SRT; Kaufman et al., 2010) – a non-verbal measure of statistical learning ability, and the noun test – a measure of rote memory. Below, we explain the tests in the order in which they were administered.

The noun test
On average, learners needed 45.72 items to pass the noun test (SD = 6.32, min = 40, max = 67). Note that on this test, lower scores indicated better performance. There were 6 learners who scored 40, which means that for those learners, the test potentially did not measure the full range of rote memory abilities. According to a Mann-Whitney U test, reliable (M = 44.6, SD = 5.11) and unreliable input learners (M = 46.84, SD = 7.26) did not differ significantly in their noun test performance (W = 260, p = .31, 95% CI [-0.4, 1]).
Chapter 2

The non-verbal IQ test

As a measure of learners’ non-verbal analytical ability, we used the Matrix Reasoning subtest of the Wechsler Adult Intelligence Scale – 4th edition (WAIS-IV) (Wechsler, 2008). The test measures non-verbal abstract problem solving and inductive reasoning ability.

The test consisted of 26 items, and the learners’ task in every item was to identify and choose – out of five options – the missing part of a visual pattern. It was administered after 15 minutes of exposure to the target structure, and learners were given maximum 20 minutes to complete the test.

Participants’ average accuracy was 22.84 (SD = 2.06, min = 17, max = 26). A Mann-Whitney U test showed that reliable (M = 23.08, SD = 2.24) and unreliable input learners (M = 22.6, SD = 1.89) did not differ significantly in their performance on this test (W = 359, p = .37, 95% CI [-1, 2]).

The serial reaction time (SRT) task

We used the probabilistic SRT task (Kaufman et al., 2010) as a visual measure of learners’ statistical learning abilities. In this task, participants responded to a visual cue – an asterisk – that appeared in one of four horizontal positions on a computer screen, while their reaction times (RTs) were recorded.

The cues followed an underlying pattern compatible either with sequence A – 85% of the time – or sequence B – 15% of the time. Sequences A and B differed in their second order information, i.e. two consecutive locations lead to a different subsequent location for these two sequences. Through repeated exposure, participants may gradually come to predict the position of the cue, which reduces their RTs on the dominant sequence A trials, and increases their RTs on sequence B trials. The difference between sequence B and sequence A reaction times provides a measure of learning, and higher scores indicate better statistical learning.

Participants were asked to respond as quickly as possible to the visual cue by pressing one of the four corresponding keys on the keyboard. They did a practice block of 48 trials and eight blocks of 120 trials each. The task lasted 8 – 10 minutes and was administered after debriefing.

We analyzed the SRT results using the procedure described in Granena (2013). For each participant, we obtained an SRT score by subtracting RTs to dominant trials from the RTs to non-dominant trials in every block, and calculating the average of all blocks. The average score was 19.17 (SD = 12.53, min = -10.04, max = 52.29), indicating that learners as a group showed statistical learning and became sensitive to the dominant pattern. A Welch two-sample t-test showed no difference between reliable (M = 20.15, SD = 13.45) and unreliable input learners (M = 18.18, SD = 11.75) in their SRT performance (t(48) = 0.55, p = .58, 95% CI [-5.21, 9.16]).
The role of unreliable input in L2 pattern learning

2.7 The LLAMA D sound recognition task
We used the LLAMA D task (Meara, 2005) as a measure of learners’ verbal working memory. The test measures the ability to recognize spoken language that you have been exposed to before. According to Meara (2005), this aptitude is involved in the learning of vocabulary and morphological patterns.

Participants first listened to a sequence of 10 words in an unknown language, after which they did a 30-item test, in which they heard both familiar and novel words, and for each word, they needed to decide whether they had heard it before or not. Whenever participants decided correctly, they gained points, and whenever they decided incorrectly, they lost points. The maximum score was 75, and the task lasted about 5 minutes. This test was administered after the SRT task.

Participants had average accuracy of 22.8 (SD = 12.78, min = 0, max = 50). According to a Welch two-sample t-test, there were no differences between the reliable (M = 22.2, SD = 9.8) and unreliable input learners (M = 23.4, SD = 15.39) in their performance on LLAMA D (t(40.71) = -0.33, p = .74, 95% CI [-8.57, 6.17]).

The LLAMA F grammatical inference task
The LLAMA F task (Meara, 2005) measured learners’ verbal analytical abilities. Participants had 5 minutes to discover grammatical patterns in an unknown language by clicking on buttons in the program. For each button, they saw an image and read a short phrase or sentence describing it. Then, they did a 20-item test in which they saw an image and read two sentences. Their task was to choose the grammatical sentence corresponding to the image. The maximum score was 100, and the task lasted about 8 minutes. It was administered after the LLAMA D task.

The average accuracy on LLAMA F was 61.8 (SD = 25.53, min = 0, max = 100). A Welch two-sample t-test showed no significant differences between reliable (M = 61.2, SD = 27.28) and unreliable input learners (M = 62.4, SD = 24.2) in their performance on LLAMA F (t(48) = -0.17, p = .87, 95% CI [-15.87, 13.47]).

2.7.7 Statistical procedures
We used R (R Development Core Team, 2011) to perform all statistical analyses. Prior to the analyses, we centered all continuous variables (e.g., aptitude measure scores) so that each learner’s individual score always reflected the difference from the mean score of all participants. For all categorical variables, we specified explicit contrasts, where we assigned -0.5 to one level of the variable and +0.5 to the other level of the variable.
To analyze learners' performance on GJTs, and the relation between aptitude measures and GJT performance, we used generalized linear mixed models (GLMMs), which include random effects in addition to fixed effects. The dependent variable was learners' accuracy. This was a binomial variable coded as 1 (correct) or 0 (incorrect). We always used maximally specified models, in which we included all fixed effects under investigation, and all two-way and three-way interactions between them.

Barr, Levy, Scheepers, and Tily (2013) recommend applying maximal random effects structures when using linear mixed models for confirmatory hypothesis testing. Following their recommendation, in our GLMMs, we always included both by-subject and by-item random intercepts and slopes for each fixed effect included in the model. Whenever subjects or items were nested within a particular fixed effect variable, we did not include the corresponding by-subject or by-item random slopes because such models would be unidentifiable. For instance, given that subjects were nested within group (i.e., reliable or unreliable group), we included the by-item but not the by-subject random slope for the group fixed effect.

When analyzing the relation between learners' aptitudes and their awareness of the target pattern, we used generalized linear models (GLMs), which include only fixed effects.

In the results section below, for each main effect or interaction, we report estimate \( (b) \), standard error \( (SE) \), probability \( (p) \), 95\% confidence intervals \( (CI) \), and the effect size in terms of odds ratios \( (OR) \), where ORs above 1 or below 1 respectively indicate an increase or a decrease in the dependent variable as a result of a particular main effect or interaction.

### 2.3 Results

Given that learners in our study showed substantial individual differences in whether or not they developed awareness of the target pattern, we took these differences into consideration in our analyses. We start this section by presenting results pertaining to learners' pattern awareness. We then look at learners' performance on the GJT with familiar nouns and their ability to generalize the pattern to novel nouns (i.e., GJT with novel nouns). Finally, we present results on how different aptitude measures predicted learners' performance on the two GJTs and their awareness of the pattern.

#### 2.3.1 Debriefing results - awareness

In the debriefing session, learners reported that they mostly focused on learning nouns, verbs, and adjectives during the exposure. Also, 36 out of 50 learners reported having awareness that lep was a masculine determiner, whereas ris was a feminine determiner. Using the information from the
debriefing, we classified learners as aware or unaware of the target pattern at different stages of the experiment. Out of 50 learners, 19 became aware of the agreement pattern during or immediately after the experiment, and could reproduce it correctly. Table 2.1 summarizes the results. As the experiment progressed, more and more learners were becoming aware of the pattern. However, most learners (n = 31) reported not having noticed the pattern at any stage.

Table 2.1. Overview of the cumulative number of learners aware of the target pattern at different stages in the experiment, split by group

<table>
<thead>
<tr>
<th>Stage in the experiment</th>
<th>Reliable input group</th>
<th>Unreliable input group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td>4/25</td>
<td>4/25</td>
</tr>
<tr>
<td>GJT novel nouns</td>
<td>8/25</td>
<td>5/25</td>
</tr>
<tr>
<td>GJT familiar nouns</td>
<td>9/25</td>
<td>7/25</td>
</tr>
<tr>
<td>After experiment</td>
<td>11/25</td>
<td>8/25</td>
</tr>
</tbody>
</table>

Learners reported that during the two GJTs they relied on memory, guessing, intuition, a wrong pattern (e.g., that one determiner denotes animate and the other inanimate objects), or they reported having judged items based on nouns only, which we labeled as wrong focus. For the analysis of the GJT results, we chose not to exclude learners who had a wrong focus or relied on a wrong pattern because such sources of reliance may have stemmed from learners’ lack of knowledge or insensitivity to the target pattern.

2.3.2 GJT with familiar nouns
We analyzed the results of the GJT with familiar nouns to find out if learners acquired the target pattern of the items they had heard in the exposure, and if their knowledge was influenced by the input type they had received (reliable vs. unreliable). Also, given that some learners were aware of the pattern during this test (n = 16; 9 in the reliable input group, 7 in the unreliable input group), we wanted to find out if input type differentially affected test performance for aware and unaware learners.

Items in this test were always scored in accordance with how they were presented during the exposure. This implies that consistent items were scored in the same way for the two learner groups, while inconsistent items or exceptions were scored differently for reliable and unreliable input learners. Therefore, we analyzed these two sets of items separately.
Consistent items analysis

First we analyzed learners’ performance on consistent items. Eight of the consistent items had biological gender, and we excluded them from the analysis because knowledge of determiner-noun agreement was not needed to respond correctly to these items; knowledge of gender only – that ris is a feminine determiner, whereas lep is a masculine determiner – could have led to the correct response on these items. A statistical analysis confirmed that learners tended to be more accurate on biological ($M = 77.00$, $SD = 24.66$) than on non-biological consistent items ($M = 68.00$, $SD = 20.04$); this difference was marginally significant, as shown by a Mann-Whitney U test ($W = 1526.5$, $p = 0.05$, 95% CI [-0.00002, 2.08]). In order to be able to draw better conclusions about learners’ knowledge of the agreement pattern, we analyzed consistent, non-biological items only ($n = 24$).

Average accuracy was 68.00% ($SD = 20.04$), and learners as a group performed significantly above chance ($b = 1.26$, $SE = 0.33$, $p < .001$, 95% CI [0.62, 1.94], $OR = 3.53$). In order to find out how input reliability affected learners’ performance, and if this differed for learners aware and unaware of the pattern, we built a model that included the following fixed effects: group (reliable vs. unreliable input), pattern awareness (whether or not learners were aware of the target pattern during the test), and grammaticality (grammatical vs. ungrammatical items). We found a main effect of group – learners in the unreliable input group were less accurate than learners in the reliable input group ($b = -1.02$, $SE = 0.41$, $p = .01$, 95% CI [-1.85, -0.2], $OR = 0.36$); a main effect of pattern awareness – learners aware of the target pattern during the test were much more accurate than learners unaware of the pattern ($b = 2.6$, $SE = 0.42$, $p < .001$, 95% CI [1.75, 3.44], $OR = 13.4$); and a main effect of grammaticality – learners were less accurate when responding to ungrammatical items, meaning that they had a bias towards marking phrases as grammatical ($b = -1.89$, $SE = 0.53$, $p < .001$, 95% CI [-2.96, -0.83], $OR = 0.15$).

There was a significant group x grammaticality interaction, i.e. unreliable input learners were less accurate than reliable input learners when responding to ungrammatical items ($b = -2.21$, $SE = 1.03$, $p = 0.03$, 95% CI [-4.28, -0.14], $OR = 0.11$), and a group x awareness interaction – when aware learners had heard unreliable input, they were less accurate than when they had heard reliable input (Figure 2.1) ($b = -1.92$, $SE = 0.82$, $p = .02$, 95% CI [-3.56, -0.28], $OR = 0.15$).

The interaction between grammaticality and awareness was not significant ($b = 1.1$, $SE = 1.05$, $p = .3$, 95% CI [-1, 3.19], $OR = 2.99$). Finally, the three-way interaction between group, pattern awareness, and grammaticality was not significant either. However, there was a tendency that aware learners were less accurate on ungrammatical items if they had received unreliable input (Table 2.2) ($b = -3.21$, $SE = 2.06$, $p = .12$, 95% CI [-7.34, 0.91], $OR = 0.04$). It is possible
that this tendency did not reach significance due to the lack of power; few learners were aware of the pattern – 9 learners in the reliable input group and 7 learners in the unreliable input group.

Given that learners aware of the pattern were disturbed by unreliable input, we were interested in whether this disturbance was also visible from learners’ debriefing reports about what they relied on during the test. In the reliable input group, out of 9 learners aware of the pattern, 8 reported having relied on it during the test. However, in the unreliable input group, out of 7 learners aware of the pattern, only 2 reportedly relied on it. This implies that learners who had received unreliable input often did not feel sufficiently confident to rely on the pattern they had noticed. However, there were substantial individual differences between aware learners in their ability to cope with input unreliability. Figure 2.1 shows accuracies of individual learners, and the horizontal lines show mean accuracies of aware and unaware learners in the reliable and unreliable input groups. As can be seen from Figure 2.1 and SDs in Table 2.2, some aware learners in the unreliable input group were strongly disturbed by unreliability of the input, while others managed to achieve accuracies comparable to those of reliable input learners.

Table 2.2. Overview of learners’ mean accuracies and their standard deviations (SD), split by group, pattern awareness, and grammaticality on the GJT with familiar nouns – consistent items

<table>
<thead>
<tr>
<th></th>
<th>Grammatical items</th>
<th>Ungrammatical items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aware learners</td>
<td>Unaware learners</td>
</tr>
<tr>
<td><strong>Group</strong></td>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
</tr>
<tr>
<td>Reliable</td>
<td>96.3 (6.05)</td>
<td>77.08 (20.53)</td>
</tr>
<tr>
<td>Unreliable</td>
<td>96.43 (4.45)</td>
<td>80.56 (17.39)</td>
</tr>
</tbody>
</table>
Considering that the analysis above could not tell us if learners unaware of the pattern performed above chance, and if their performance was affected by the input type, we ran a separate analysis on learners who were unaware of the target pattern during the test \((n = 34, M = 58.95, SD = 13.61)\). We found a trend towards overall above chance performance \((b = 0.48, SE = 0.26, p = .06, 95\% \text{ CI } [-0.04, 1.01], \ OR = 1.61)\), which together with large standard deviations (Table 2.2 above) indicates that there were many individual differences in this group. We then built a model that included group and grammaticality as fixed effects. The analysis showed no main effect of group \((b = -0.04, SE = 0.28, p = .88, 95\% \text{ CI } [-0.6, 0.51], \ OR = 0.96)\), but there was a main effect of grammaticality \((b = -2.42, SE = 0.51, p < .001, 95\% \text{ CI } [-3.43, -1.4], \ OR = 0.09)\). The group x grammaticality interaction was not significant \((b = -0.58, SE = 0.92, p = .53, 95\% \text{ CI } [-2.41, 1.25], \ OR = 0.56)\).

In order to identify individual unaware learners who performed reliably above chance, we used a binomial test, which tests if a score is above chance given the number of trials. Learners who scored 75\% or above could be classified as having performed above chance \((p < .05)\). We found 4 learners who were unaware of the pattern during the test, but performed reliably above chance. Three learners were in the reliable input group, and 1 was in the unreliable input group. Two of the 4 learners reported having relied on intuition during the test, whereas the other 2 relied on memory.

To sum up, the results suggest that unreliable input had a negative impact on learners’ test performance when they were aware of the pattern. Learners
often did not feel confident enough to rely on the pattern they had noticed, and they were less accurate. There is also evidence that some learners acquired knowledge in the absence of pattern awareness, but we found no evidence that this learning was influenced by input reliability.

**Inconsistent items analysis**

Since the GJT with familiar nouns featured nouns that learners had heard in the exposure, we wanted to check if learners’ performance on the test was the result of pure item-based knowledge, or also at least partially reflected more abstract pattern knowledge. To answer this question, we analyzed the inconsistent items or exceptions. Accuracy on these items for each of the two groups reflected what they had heard during the exposure. If unreliable input learners are less accurate on inconsistent items than reliable input learners, this may be an indication that they did not just memorize items from the input, but that they overgeneralized the predominant pattern to the exception items, which would be the evidence of pattern learning (Tomasello, 2000). Reliable input learners had an average accuracy of 77.5% ($SD = 18.75$) on the inconsistent items, whereas the average accuracy of unreliable input learners was 53% ($SD = 20.5$).

Overall, the participants performed significantly above chance ($b = 0.72, SE = 0.17, p < .001, 95\% CI [0.39, 1.1], OR = 2.06$). In order to find out if learners overgeneralized the dominant pattern to exceptions, and if this differed for aware and unaware learners, we built a model that included the following fixed effects: group, pattern awareness, and grammaticality. There were main effects of group – unreliable input learners were less accurate ($b = -2.03, SE = 0.48, p < .001, 95\% CI [-3, -1.07], OR = 0.13$); pattern awareness – learners aware of the pattern were more accurate ($b = 1.01, SE = 0.45, p = .026, 95\% CI [0.1, 1.91], OR = 2.73$); and grammaticality – learners were less accurate on ungrammatical items ($b = -1.84, SE = 0.53, p < .001, 95\% CI [-2.89, -0.78], OR = 0.16$). The interaction between group and pattern awareness was also significant (Table 2.3; Figure 2.2), meaning that learners who had heard unreliable input made more mistakes on inconsistent items if they were aware of the pattern ($b = -2.47, SE = 0.9, p = .006, 95\% CI [-4.26, -0.68], OR = 0.08$). The other two-way interactions were not significant: group x grammaticality ($b = -0.75, SE = 1, p = .45, 95\% CI [-2.75, 1.24], OR = 0.47$) and awareness x grammaticality ($b = -0.28, SE = 1, p = .78, 95\% CI [-2.27, 1.72], OR = 0.76$). The three-way interaction between group, awareness, and grammaticality was not significant either ($b = -0.02, SE = 2.01, p = .99, 95\% CI [-4.04, 3.99], OR = 0.98$). The fact that aware unreliable input learners made more mistakes on exceptions than aware reliable input learners indicates that they overgeneralized the pattern to the exceptions. This confirms that aware
learners did not purely memorize the items they had heard during the exposure, but that they acquired the pattern, which is in line with the pattern awareness they reported later during debriefing.

**Table 2.3. Overview of learners’ mean accuracies and their standard deviations (SD), split by group, pattern awareness, and grammaticality on the GJT with familiar nouns – inconsistent items**

<table>
<thead>
<tr>
<th></th>
<th>Grammatical items</th>
<th>Ungrammatical items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aware learners</td>
<td>Unaware learners</td>
</tr>
<tr>
<td>Group</td>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
</tr>
<tr>
<td>Reliable</td>
<td>96.67 (5.59)</td>
<td>80.00 (18.26)</td>
</tr>
<tr>
<td>Unreliable</td>
<td>91.43 (7.48)</td>
<td>80.55 (16.97)</td>
</tr>
</tbody>
</table>

**Figure 2.2. Overview of accuracies on the GJT with familiar nouns – inconsistent items, for each individual learner, split by group and pattern awareness**

Next, we wanted to check if learners also overgeneralized the pattern to exceptions and showed pattern learning when they were unaware of the pattern. We ran an analysis on unaware learners only, with group and grammaticality as fixed effects in the model. There was a main effect of group
The role of unreliable input in L2 pattern learning (Table 2.3; Figure 2.2), such that the unreliable input group was less accurate on inconsistent items than the reliable input group \( (b = -0.8, SE = 0.38, p = .03, 95\% \text{ CI } [-1.55, -0.05], OR = 0.45) \). The effect of grammaticality was also significant \( (b = -1.58, SE = 0.5, p = .002, 95\% \text{ CI } [-2.58, -0.58], OR = 0.21) \). The interaction between group and grammaticality was not significant \( (b = -0.69, SE = 0.88, p = .44, 95\% \text{ CI } [-2.45, 1.07], OR = 0.5) \). This suggests that even learners who reported being unaware of the pattern during the task overgeneralized the dominant pattern to exceptions, which may be taken as evidence of implicit pattern learning. If learners had only memorized items from the input, there would have been no reason for unreliable input learners to perform worse on exceptions than reliable input learners, because the accuracy score for both groups reflected what they had heard during the exposure.

2.3.3 GJT with novel nouns

The goal of this task was to find out if learners could generalize the target pattern to novel nouns, and if this ability was affected by the input type they had received. Also, given that some learners were aware of the pattern during this test \( (n = 13; 8 \text{ in the reliable input group, 5 in the unreliable input group}) \), whereas the others were not \( (n = 37) \), we wanted to find out whether aware and unaware learners were differentially affected by input type.

The average accuracy on the test was 60.15% \( (SD = 19.3) \). Learners as a group performed significantly above chance \( (b = 0.68, SE = 0.23, p = .003, 95\% \text{ CI } [0.23, 1.14], OR = 1.97) \). We built a model that included the following fixed effects: group, awareness, and grammaticality. There was a main effect of group – unreliable input learners were less accurate than reliable input learners \( (b = -1.16, SE = 0.28, p < .001, 95\% \text{ CI } [-1.72, -0.59], OR = 0.31) \); a main effect of pattern awareness – learners aware of the pattern during the test were more accurate \( (b = 2.55, SE = 0.28, p < .001, 95\% \text{ CI } [1.99, 3.12], OR = 12.86) \); and a main effect of grammaticality – learners were less accurate on ungrammatical than on grammatical items \( (b = -1.44, SE = 0.41, p < .001, 95\% \text{ CI } [-2.26, -0.61], OR = 0.24) \). We also found a significant group x pattern awareness interaction (Table 2.4; Figure 2.3), such that when learners were aware of the target pattern, they were less accurate if they had received unreliable input \( (b = -2.34, SE = 0.57, p < .001, 95\% \text{ CI } [-3.47, -1.21], OR = 0.1) \). The other two-way interactions were not significant: group x grammaticality \( (b = -0.25, SE = 0.79, p = .75, 95\% \text{ CI } [-1.82, 1.33], OR = 0.78) \) and awareness x grammaticality \( (b = 0.2, SE = 0.79, p = .8, 95\% \text{ CI } [-1.37, 1.78], OR = 1.23) \). The three-way interaction between group, grammaticality, and awareness was not significant either \( (b = -0.08, SE = 1.57, p = .96, 95\% \text{ CI } [-3.23, 3.07], OR = 0.92) \).
We next looked at aware learners’ debriefing reports to see whether they relied on the pattern during this test, and whether this differed in the reliable and unreliable input group. We found that all 8 reliable input learners reported having relied on the pattern during the test. However, out of 5 unreliable input learners aware of the pattern, 3 reported having relied on it.

**Table 2.4. Overview of learners’ mean accuracies and their standard deviations (SD), split by group, pattern awareness, and grammaticality on the GJT with novel nouns**

<table>
<thead>
<tr>
<th></th>
<th>Grammatical items</th>
<th>Ungrammatical items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aware learners</td>
<td>Unaware learners</td>
</tr>
<tr>
<td><strong>Group</strong></td>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
</tr>
<tr>
<td>Reliable</td>
<td>98.12 (2.59)</td>
<td>65.29 (14.63)</td>
</tr>
<tr>
<td>Unreliable</td>
<td>87.00 (15.65)</td>
<td>67.00 (18.09)</td>
</tr>
</tbody>
</table>

**Figure 2.3. Overview of accuracies on the GJT novel nouns, for each individual learner, split by group and pattern awareness**

The results show that learners aware of the target pattern could generalize the pattern they had noticed to novel items. Their ability to do so was negatively affected by the presence of exceptions in the input, but again, there
were substantial individual differences, and some aware learners had less difficulty coping with unreliable input than others (Figure 2.3).

The statistical analysis presented above could not tell us if learners unaware of the pattern could generalize the pattern above chance, or whether they were negatively affected by unreliable input. Therefore, we ran a separate analysis on accuracies of unaware learners only ($M = 51.28, SD = 8.61$). We found no above-chance performance ($b = 0.06, SE = 0.13, p = 0.65, 95\% CI [-0.2, 0.32], OR = 1.06$). There was no main effect of group ($b = 0.03, SE = 0.14, p = .84, 95\% CI [-0.25, 0.3], OR = 1.03$), but there was a main effect of grammaticality – learners were less accurate on ungrammatical items ($b = -1.47, SE = 0.31, p < .001, 95\% CI [-2.09, -0.86], OR = 0.23$). The interaction between group and grammaticality was not significant ($b = -0.19, SE = 0.58, p = .75, 95\% CI [-1.35, 0.98], OR = 0.83$).

Finally, we wanted to check if there were any learners unaware of the pattern who performed significantly above chance on the test. According to the binomial test, learners who scored 67.5% or more could be reliably classified as having performed above chance ($p < .05$). There were 2 such learners. One was in the reliable input group, had accuracy of 70%, and claimed to have guessed, whereas the other was in the unreliable input group, had 80% accuracy, and used intuition.

The results of the GJT with novel nouns show that when learners were aware of the pattern, the unreliable input they had been exposed to often disturbed their ability to generalize the pattern to new nouns. Unaware learners as a group did not show above chance performance, although there may have been two learners who generalized the pattern to new nouns in the absence of pattern awareness.

2.3.4 The role of working memory, rote memory, analytical ability, and statistical learning ability

We were interested in whether the aptitude measures used in this study predicted learners’ performance on the two GJT’s, and whether this differed for the two input conditions. Finally, we looked at whether any aptitude measures could predict if learners would become aware of the pattern. Given that we were interested in whether each of the aptitude measures could predict learning, we ran separate models for all aptitude measures under investigation.

First, we checked correlations between the different aptitude measures we used, and we only found a significant positive correlation between LLAMA F and IQ scores ($r = 0.35, p = .01$), both tapping into learners’ analytical abilities.

As Table 2.5 and Figure 2.4 show, LLAMA D, noun test scores, and IQ scores significantly predicted learners’ performance on the GJT with familiar nouns – non-biological, consistent items. These results suggest that learners who
performed better on measures of working memory, rote memory, and analytical ability were also more accurate on the GJT with familiar nouns.

Table 2.5. Overview of aptitude involvement in learners’ performance on the GJT with familiar nouns, GJT with novel nouns, and learners’ pattern awareness

<table>
<thead>
<tr>
<th>GJT WITH FAMILIAR NOUNS</th>
<th>b</th>
<th>SE</th>
<th>p</th>
<th>95% CI</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main effects of aptitude measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D</td>
<td>0.06</td>
<td>0.02</td>
<td>&lt;.001</td>
<td>[0.02, 0.09]</td>
<td>1.06</td>
</tr>
<tr>
<td>Noun test</td>
<td>-0.05</td>
<td>0.02</td>
<td>.006</td>
<td>[-0.08, -0.01]</td>
<td>0.95</td>
</tr>
<tr>
<td>IQ</td>
<td>0.17</td>
<td>0.08</td>
<td>.03</td>
<td>[0.01, 0.34]</td>
<td>1.19</td>
</tr>
<tr>
<td>LLAMA F</td>
<td>0.01</td>
<td>0.01</td>
<td>.07</td>
<td>[-0.001, 0.02]</td>
<td>0.01</td>
</tr>
<tr>
<td>SRT</td>
<td>-0.01</td>
<td>0.02</td>
<td>.55</td>
<td>[-0.06, 0.03]</td>
<td>0.99</td>
</tr>
<tr>
<td><strong>Aptitude x group interactions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D x group</td>
<td>-0.04</td>
<td>0.03</td>
<td>.25</td>
<td>[-0.11, 0.03]</td>
<td>0.96</td>
</tr>
<tr>
<td>Noun test x group</td>
<td>-0.09</td>
<td>0.09</td>
<td>.32</td>
<td>[-0.28, 0.1]</td>
<td>0.91</td>
</tr>
<tr>
<td>IQ x group</td>
<td>-0.02</td>
<td>0.15</td>
<td>.91</td>
<td>[-0.31, 0.28]</td>
<td>0.98</td>
</tr>
<tr>
<td>LLAMA F x group</td>
<td>-0.01</td>
<td>0.01</td>
<td>.59</td>
<td>[-0.03, 0.02]</td>
<td>0.99</td>
</tr>
<tr>
<td>SRT x group</td>
<td>0.01</td>
<td>0.04</td>
<td>.8</td>
<td>[-0.07, 0.09]</td>
<td>1.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GJT WITH NOVEL NOUNS</th>
<th>b</th>
<th>SE</th>
<th>p</th>
<th>95% CI</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main effects of aptitude measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D</td>
<td>0.02</td>
<td>0.01</td>
<td>.19</td>
<td>[-0.01, 0.05]</td>
<td>1.02</td>
</tr>
<tr>
<td>Noun test</td>
<td>-0.07</td>
<td>0.04</td>
<td>.1</td>
<td>[-0.16, 0.02]</td>
<td>0.93</td>
</tr>
<tr>
<td>IQ</td>
<td>0.13</td>
<td>0.13</td>
<td>.34</td>
<td>[-0.14, 0.4]</td>
<td>1.14</td>
</tr>
<tr>
<td>LLAMA F</td>
<td>0.02</td>
<td>0.01</td>
<td>.06</td>
<td>[-0.003, 0.03]</td>
<td>1.01</td>
</tr>
<tr>
<td>SRT</td>
<td>0.002</td>
<td>0.03</td>
<td>.95</td>
<td>[-0.05, 0.05]</td>
<td>1.002</td>
</tr>
<tr>
<td><strong>Aptitude x group interactions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D x group</td>
<td>-0.08</td>
<td>0.03</td>
<td>.02</td>
<td>[-0.15, -0.01]</td>
<td>0.92</td>
</tr>
<tr>
<td>Noun test x group</td>
<td>-0.06</td>
<td>0.08</td>
<td>.44</td>
<td>[-0.22, 0.1]</td>
<td>0.94</td>
</tr>
<tr>
<td>IQ x group</td>
<td>-0.08</td>
<td>0.27</td>
<td>.77</td>
<td>[-0.61, 0.45]</td>
<td>0.92</td>
</tr>
<tr>
<td>LLAMA F x group</td>
<td>-0.01</td>
<td>0.02</td>
<td>.66</td>
<td>[-0.05, 0.03]</td>
<td>0.99</td>
</tr>
<tr>
<td>SRT x group</td>
<td>-0.004</td>
<td>0.05</td>
<td>.93</td>
<td>[-0.1, 0.09]</td>
<td>0.996</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PATTERN AWARENESS</th>
<th>b</th>
<th>SE</th>
<th>p</th>
<th>95% CI</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLAMA D</td>
<td>0.05</td>
<td>0.025</td>
<td>.03</td>
<td>[0.01, 0.11]</td>
<td>1.06</td>
</tr>
<tr>
<td>Noun test</td>
<td>-0.08</td>
<td>0.06</td>
<td>.17</td>
<td>[-0.2, 0.04]</td>
<td>0.92</td>
</tr>
<tr>
<td>IQ</td>
<td>0.48</td>
<td>0.19</td>
<td>.01</td>
<td>[0.14, 0.89]</td>
<td>1.62</td>
</tr>
<tr>
<td>LLAMA F</td>
<td>0.03</td>
<td>0.01</td>
<td>.07</td>
<td>[0.0003, 0.05]</td>
<td>1.03</td>
</tr>
<tr>
<td>SRT</td>
<td>0.02</td>
<td>0.02</td>
<td>.51</td>
<td>[-0.03, 0.06]</td>
<td>1.02</td>
</tr>
</tbody>
</table>
The role of unreliable input in L2 pattern learning

Figure 2.4. Relation between learners’ predicted performance on the GJT with familiar nouns and a) the LLAMA D scores, b) the noun test scores, and c) the IQ scores.

The involvement of none of the aptitude measures differed for the two input groups, i.e. there were no significant aptitude by treatment interactions (Table 2.5).

When it comes to the GJT with novel nouns, none of the aptitude tests significantly predicted learners’ performance on that test, although there were some predictive tendencies with LLAMA F scores and noun test scores (Table 2.5). When it comes to aptitude by group interactions, we found an interaction between group and LLAMA D (Table 2.5; Figure 2.5), such that LLAMA D scores predicted performance of reliable input learners, but not performance of unreliable input learners.
Finally, we checked whether any of the aptitude measures predicted if learners would develop awareness of the target pattern during or after the experiment. We found that both LLAMA D and IQ predicted learners’ pattern awareness (Table 2.5), which implies that learners who had better working memory and analytical ability were more likely to develop pattern awareness.

2.4 Discussion

The first and main goal of this study was to investigate if input reliability affects pattern learning in SLA, which may be expected based on previous L1 research (e.g., Costa et al., 2015; Miller, 2007; Miller & Schmitt, 2012). We provide evidence that unreliable input with exceptions negatively affects pattern learning in adult SLA. Learners’ performance on both GJTs was negatively affected by unreliable input, but this negative effect was present only when learners were aware of the pattern. This may imply that exceptions in the input disturbed learners’ mental representations of the pattern; learners were less confident in using the pattern they had noticed during the test, which resulted in a lower accuracy. These findings are in line with L1 research results that also suggest that inconsistent input negatively affects the acquisition of linguistic structures (e.g., Costa et al., 2015; Miller, 2007; Miller & Schmitt, 2012). However, it should be pointed out that while L1 acquisition is considered to be mostly implicit, L2 acquisition is often thought to be mediated by awareness. We found that aware learners were negatively affected by exceptions in the input, but it is important to bear in mind that there were substantial individual differences. While some aware learners were obviously disturbed by exceptions, others managed to achieve accuracies comparable to reliable input learners. When it comes to implications for L2 teaching, the results may suggest that – if practically possible – offering exceptions later in the learning process...
may be more conducive to pattern learning than offering them straight from the beginning, when the pattern has not been fully established and entrenched yet. Alternatively, teachers could try to reduce the ratio of exceptions to regular items in the initial stages of pattern learning.

The second goal of the study was to explore whether pattern learning could be predicted by several aptitude measures: working memory, rote memory, analytical ability, and statistical learning ability. We found several links between learning and memory capacities. LLAMA D and noun test scores predicted learners’ performance on the GJT that featured familiar nouns. Considering that both LLAMA D and noun test scores measured learners’ sensitivity to previously heard linguistic input, it is not surprising that these measures were involved in learners’ performance on the GJT that allowed for the use of item-based knowledge. Importantly, LLAMA D was also involved in learners’ ability to generalize patterns to novel nouns, but this was the case for reliable input learners only. The lack of LLAMA D involvement in pattern learning of unreliable input learners may be due to the fact that even learners who had high working memory ability did not benefit from it because they were confused by exceptions in the input. The links with LLAMA D that we observed are in line with Granena (2013), who also found that LLAMA D was involved in the acquisition of agreement structures. When it comes to the links with analytical ability, we found that IQ test scores predicted learners’ performance on the GJT with familiar nouns but not on the GJT with novel nouns. Unlike some previous studies (Brooks & Kempe, 2013; Granena, 2013; McDonough & Trofimovich, 2016), we found no links between pattern learning and statistical learning ability in this study. This may be due to the fact that the SRT task measures implicit statistical learning ability, and in our sample much of the learning was driven or accompanied by awareness.

In this study, we also observed links between aptitudes and learners’ awareness of the target pattern, i.e. learners who scored higher on working memory (LLAMA D) and analytical ability (IQ) were more likely to also develop awareness of the target pattern. The finding that working memory was involved in developing awareness of the target pattern is in line with previous research that found links between working memory and explicit but not implicit learning (e.g., Tagarelli et al., 2011; Unsworth & Engle, 2005; Yang & Li, 2012).

When it comes to the question of aptitude by treatment interactions (ATIs), i.e. whether involvement of different aptitude measures differed for reliable and unreliable input types, we found an interaction between working memory and input reliability. Our initial hypothesis was that working memory may be more involved in learning from unreliable input than in learning from categorical input given that unreliable input may require memorizing regular items and exceptions. However, we found the opposite result. Namely, LLAMA
D predicted learners’ performance on the GJT with novel nouns for the reliable input learners but not for the unreliable input learners. As pointed out before, it is possible that beneficial effects of working memory on learning from unreliable input were overridden by the disturbing effects of unreliable input. The aptitude by treatment interaction that we found in this study confirms the complex picture of the involvement of aptitude in learning (DeKeyser, 2012; Robinson, 2002, 2005a; Skehan, 2002), and calls for more research that looks beyond simple correlations between aptitudes and learning outcomes.

In our study, we found evidence of both item-based and abstract pattern knowledge. Learners’ above chance performance on the GJT with novel nouns indicates that at least some learners acquired abstract, generalizable knowledge of the pattern. Also, in the GJT with familiar nouns, learners in the unreliable condition showed clear overgeneralizations of the dominant pattern to exceptions, which also points to the presence of more abstract pattern knowledge (Tomasello, 2000). There is also evidence of existence of item learning. First, learners were less accurate on the GJT with novel nouns than on the GJT with familiar nouns, which allowed both application of item-based and pattern-based knowledge. Second, some learners reported having relied on memory during the GJT with familiar nouns. It is possible that pattern learning in our study followed the trajectory suggested by usage-based approaches, i.e. from item-based learning to generalizing the pattern (Bybee, 2008; Goldberg, 2006; N. C. Ellis, 2002; Tomasello, 2000). This process was often mediated and speeded up by awareness.

Although we did not set out to investigate implicit learning, we provide some evidence of the acquisition of pattern knowledge in the absence of awareness (in line with Rebuschat et al., 2015; Williams, 2005). Overgeneralization of the pattern to exceptions was even found in learners who were unaware of the pattern during the task. This may prove that they acquired abstract pattern knowledge (Tomasello, 2000). Generalizing the pattern to novel items in the absence of awareness was much more difficult, and only a few learners showed tendencies in this direction. This study offers a potentially fruitful paradigm for investigating early stages of implicit learning: a paradigm that does not necessarily require learners to generalize a pattern to novel items, which learners often find unpleasant and confusing, but allows them to demonstrate implicit pattern learning on already encountered exception items.

The overall results of the present study confirm that adult L2 learners differ considerably in their language learning success, even when context variables, such as amount and kind of input, are identical for them. This is quite striking given that learners in our study were highly educated adults, who could be considered a relatively homogeneous group, likely to be highly successful. Some learners developed awareness of the target pattern, whereas the others did not. Also, for learners who did become aware of the pattern, the awareness
developed at different moments during the experiment. The results imply that individual differences are more present than normally assumed in research that makes group comparisons. This has deep implications for the predominant focus on group results in the literature, and shows a clear need for more research focusing on individual differences (Dabrowska & Street, 2006).

The study also illustrates the importance of using debriefing or retrospective verbal reports in experimental research, which can complement the test results and help understand them better (Rebuschat et al., 2015). In our case, crucial information about learners’ awareness was gained from the debriefing session.

Finally, several limitations of the study need to be pointed out. The first set of limitations pertains to the aptitude-related results of this study. The links with aptitudes that we find need to be treated with caution given that most of the effects we found are quite small. Also, a question that remains open is whether the observed relations between aptitudes and learning success were causal. Direct evidence for causality can only come from studies that assign learners to particular treatment groups based on their aptitude profiles, and then find a relation between aptitude and learning (e.g., Wesche, 1981). Next, it needs to be pointed out that the exposure to the language and the target structure in this study was very brief, so only first signs of learning could be observed. It would be important to investigate reliability and aptitude effects using longer exposure and delayed posttests. The use of retrospective verbal reports as a way of establishing if learners became aware of the pattern could be seen as another limitation of this study. There has been some discussion about the validity of this awareness measure because retrospective verbal reports may lead to overestimating the number of unaware learners (Rebuschat et al., 2015). However, other methods of distinguishing between aware and unaware learners are not without drawbacks. As noted in Rebuschat et al. (2015), think-aloud protocols often fail to detect aware learners, whereas trial-by-trial source attributions (Dienes & Scott, 2005) clearly invite learners to search for the rules by suggesting rule as one of the options to choose from. For an extensive discussion of advantages and disadvantages of different awareness measures, see Rebuschat et al. (2015). Another limitation of the study is that our evidence of implicit learning is based on a GJT, which is certainly not the best task for measuring implicit knowledge, especially because it can be reactive in the sense that it can trigger learners’ search for patterns (Rebuschat et al., 2015). The use of online techniques that do not favor the application of explicit knowledge, and are less likely to trigger awareness would be more appropriate for drawing conclusions on implicit learning (Andringa & Rebuschat, 2015). One such measure is visual world eye-tracking (e.g., Andringa & Curcic, 2015; Wonnacott, Newport, & Tanenhaus, 2008), where learners are never exposed to ungrammatical stimuli, unlike in GJTs and
even in many online techniques designed to measure implicit knowledge, such as event-related potentials (ERPs) and self-paced listening.
Chapter 3: Effects of prolonged exposure on L2 pattern learning from reliable and unreliable input

Abstract

This study uses prolonged exposure to try and replicate the Chapter 2 findings related to the effects of input reliability, working memory, and analytical ability in learning of determiner-noun agreement pattern in adults. The motivation for this partial replication was also to find out whether prolonged exposure to the target pattern would lead to higher rates of pattern awareness among learners, or to higher levels of implicit pattern learning. The results of this study fully replicated the previously observed disturbing effects of unreliable input among learners aware of the target pattern. However, when learners were unaware of the pattern and received longer exposure to the target structure, unreliable input helped them to better memorize both regular items and exceptions. Unlike in the short exposure study, we found no evidence of implicit pattern learning. We observed that learners who had received longer exposure were more likely to develop awareness of the pattern compared to learners who had received short exposure. However, this effect turned out to be driven by differences in working memory between learners in the short and long exposure groups. The present study partially replicates the relations between aptitudes and pattern learning observed in the initial Chapter 2 study.

3.1 Introduction

Adult second language learners differ substantially in their ability to extract patterns from the language input they are exposed to. The process of pattern learning and language learning in general is influenced among other factors by the characteristics of the language input (Bybee, 2008; N. C. Ellis, 2002) and by learners’ cognitive abilities or aptitudes (Dörnyei, 2006). In order to find out how particular input characteristics facilitate or disturb L2 pattern learning, several studies have experimentally modified features of the input and looked at how such modifications affect learning (e.g., Casenhiser & Goldberg, 2005; McDonough & Nekrasova-Becker, 2014; Year & Gordon, 2009).

Input reliability or consistency has been identified as an input feature that may influence L2 pattern learning (e.g., DeKeyser, 1994; Hulstijn & De Graaff, 1994). Several L1 studies have found experimental evidence that unreliable input can delay child acquisition of certain structures (e.g., Costa et al., 2015; Miller, 2007; Miller & Schmitt, 2012). However, there has been almost no experimental research on this topic in L2 learning. In Chapter 2, we
investigated how input reliability affects L2 pattern learning, and we found that exceptions considerably disturbed learning, but only when learners had already become aware of the target pattern. We also found evidence of implicit pattern learning without awareness, but there was no evidence that this learning was disturbed by input unreliability.

The main goal of the present study was to investigate whether prolonged exposure to the pattern would lead to higher levels of implicit pattern learning or to higher rates of pattern awareness among learners, or both, which would yield important insights into how pattern learning progresses. For this purpose, we compared the implicit learning results and the awareness rates in the present study with those reported in the initial study presented in Chapter 2. Additionally, we wanted to check if we would replicate the findings reported in Chapter 2 related to the effects of input reliability and involvement of aptitude measures in pattern learning.

### 3.1.1 Role of input reliability in L2 pattern learning

The relation between L2 pattern learning and input characteristics, such as complexity, frequency, skewed vs. balanced input, has received some attention in experimental research (e.g., De Graaff, 1997; Denhovska et al., 2015; Goldberg et al., 2004; Housen et al., 2005; McDonough & Nekrasova-Becker, 2014; Spada & Tomita, 2010; Year & Gordon, 2009; for a review, see Chapter 2). Although the literature recognizes input reliability as a feature that can influence pattern learning success (e.g., Hulstijn & De Graaff, 1994), this feature has received almost no attention in experimental L2 research. Several L1 studies have found evidence that unreliable input delays L1 acquisition of structures such as plural marking and clitic placement (e.g., Costa et al., 2015; Miller, 2007; Miller & Schmitt, 2012). For a more detailed discussion of the L1 literature on the effects of input reliability, see Chapter 2.

A distinction between categorical, reliable rules and probabilistic, unreliable rules has been recognized in L2 acquisition research (e.g., DeKeyser, 1994; Hulstijn & De Graaff, 1994). It has also been hypothesized that reliability of rules may lead to differences in their learning (e.g., Hulstijn & De Graaff, 1994). To our knowledge, the study reported in Chapter 2 is the only study that experimentally investigated how input reliability affects learning of patterns in adult L2 learners. We will here briefly summarize the design and results of that study given that the present study aims to replicate its findings and provide more evidence of the effects observed.

The study reported in Chapter 2 was conducted with 50 native speakers of Dutch, who were exposed to a miniature language based on Fijian. The target structure was determiner-noun agreement: masculine determiner *lep* preceded nouns ending in *–uk* (*lep ganeuk*), whereas feminine determiner *ris* preceded
Effects of prolonged exposure on L2 pattern learning

nouns ending in –is (ris maramis). The reliable input group received input that contained the target pattern without any exceptions, whereas for the unreliable input group, 4 out of 20 nouns in the input were exceptions: lep preceded nouns ending in –is, and ris preceded nouns ending in –uk. Learners’ knowledge of the target pattern was measured by means of two oral grammaticality judgment tasks (GJTs) – one with novel and one with familiar nouns. The results of the study showed that learners who had heard input with exceptions did not perform as well on the GJTs as learners who had been exposed to entirely regular input. However, this was the case only when learners were aware of the target pattern, as indicated by their reports during debriefing. There was no evidence that input reliability influenced performance of learners who were unaware of the pattern. The study also found some evidence of implicit learning without pattern awareness, and this evidence came from learners’ grammaticality judgments on exceptions: learners unaware of the pattern judged exceptions according to the pattern, rather than according to what they had heard in the input, which means that they overgeneralized the dominant pattern to the exceptions. The study also found links between learners’ pattern learning and awareness on the one hand, and their analytical and working memory capacities on the other hand. In addition, we found an interaction between working memory and group, such that working memory was predictive of pattern learning from reliable but not unreliable input. In the present study, we wanted to replicate the effects of unreliable input that we had observed in the initial study, as well as the relation between pattern learning and aptitudes that we had found.

3.1.2 Implicit learning and awareness – how they interact and develop over time
There are substantial individual differences in learners’ ability to learn patterns during language exposure. Several studies have shown that even when receiving identical exposure to a novel language, some learners develop implicit knowledge, while others (also) develop explicit knowledge and awareness of the underlying patterns (e.g., Batterink, Reber, Neville, & Paller, 2015; Williams, 2005; the study reported in Chapter 2). When implicit learning effects are observed, they are typically quite small – slightly above chance performance (e.g., Williams, 2005), or they are not very robust and are restricted to certain tasks (e.g., Batterink et al., 2015; the study reported in Chapter 2 – found in GJTs with familiar, but not novel nouns). In any case, there is evidence that language exposure can trigger the development of both implicit knowledge and awareness of the pattern, although it still remains unclear how implicit knowledge and awareness develop over time, whether and how they interact, and whether accumulated implicit knowledge may foster awareness.
The relation between implicit and explicit knowledge has received much theoretical attention in the interface debate, where the focus was mostly on how explicit knowledge influences implicit knowledge. This debate includes several different positions. According to the no-interface position (Krashen, 1981, 1985, 1994) implicit and explicit knowledge do not interact because they result from different learning processes: implicit knowledge results from implicit learning, whereas explicit knowledge results from explicit learning processes, and cannot contribute to implicit knowledge. Other positions suggest that implicit and explicit knowledge do interact. The strong-interface position assumes that explicit knowledge can convert into implicit knowledge (DeKeyser, 1998, 2007; Hulstijn, 1999; O’Malley, Chamot, & Walker, 1987; Schmidt, 1995), whereas the weak-interface position suggests that explicit knowledge can indirectly foster implicit learning and knowledge (Doughty & Williams, 1998; N. C. Ellis, 2005; R. Ellis, 1997).

There has been little attention in the literature for interaction in the opposite direction, i.e. whether and how implicit knowledge affects explicit knowledge and awareness, and whether learners develop awareness as a result of accumulated implicit knowledge. According to Bialystok (1989, 1994), language learning is as a process of developing awareness of the L2 structures through analysis. By this, she means that mental representations (i.e., implicit knowledge) may change into explicit representations (i.e., explicit knowledge) through analysis. According to Bialystok, analysis leads to an increase in accessibility of knowledge, which implies that if learners spend more effort analyzing the input, they will develop more awareness. On the other hand, Cleeremans (2011) proposes that development of awareness depends on the strength of the implicit representations, rather than the efforts invested in analysis. He proposes the following principle regarding the interaction of implicit representations and awareness: “Availability to consciousness depends on the quality of representation, where quality of representation is a graded dimension defined over stability in time, strength, and distinctiveness” (p. 6). In this view, we begin with weak representations characteristic of implicit cognition. As training continues and learning progresses, these representations become stronger and more accessible for consciousness. This view implies that implicit knowledge accumulated over time may lead to the development of awareness once implicit representations have gained sufficient strength to become accessible to awareness.

The effects of exposure length on the development of implicit knowledge and awareness have not been systematically investigated. Safran et al. (1997) investigated incidental pattern learning in the artificial language learning paradigm, and found that prolonged exposure led to enhanced pattern learning, as indicated by learners’ higher accuracy rates on an auditory forced-choice recognition task. However, even though they noted that some of the learners in
their experiments developed awareness, they did not try to tease apart what kind of knowledge improved with longer exposure, i.e. whether better performance was based on implicit knowledge or explicit knowledge with awareness, or both.

In the present study we aimed to investigate the relationship between implicit learning and awareness, by comparing how implicit learning and awareness develop after shorter vs. longer exposure to the L2 pattern. Our goal was to see if longer exposure would lead to higher levels of implicit pattern learning or to more learners developing pattern awareness, or both.

### 3.1.3 The present study

Participants in this study received exposure to a miniature language based on Fijian and the determiner-noun agreement pattern. The exposure was identical to the one used in the initial study reported in Chapter 2, with the only difference being that the exposure consisted of 60% more trials. As in the initial study, learners’ target structure knowledge was tested using two oral grammaticality judgment tasks (GJTs): one with items that learners had heard during the exposure, and one with novel items in order to see if learners could generalize the pattern to the items they had not heard before.

We administered the aptitude measures that predicted learning in the initial study: the LLAMA D sound recognition task (Meara, 2005), tapping into learners’ verbal working memory and a non-verbal IQ test (Wechsler, 2008), tapping into analytical ability. We omitted the serial reaction time task (SRT; Kaufman et al., 2010) and the LLAMA F task (Meara, 2005), but added another, non-verbal measure of working memory: the visual digit span task (Wechsler, 1997). Our goal was to answer the following research questions:

1) **Does unreliable input with exceptions disturb pattern learning in the second language?** (replication of Chapter 2 results)

2) **Can analytical ability and working memory predict pattern learning and awareness? Are there aptitude by group (reliable vs. unreliable input) interactions?** (replication of Chapter 2 results)

3) **Does prolonged exposure to the target pattern lead to higher levels of implicit learning, to more awareness of the target pattern, or to both?**

We also collected eye-tracking data to investigate individual differences in learners’ online processing of the target structure; these results are presented in Chapter 5.
3.2 Method

3.2.1 Participants
Participants were 50 adult native speakers of Dutch, between 19 and 33 years old ($M_{age} = 22.56$, $SD = 3.23$). They were students or graduates majoring in different fields, but they received no education in linguistics. Participants reported having good hearing, normal or corrected-to-normal vision, no history of dyslexia, and no color-blindness. They were randomly assigned to one of the two experimental groups: the reliable input group (7 males, 18 females, $M_{age} = 22.00$, $SD = 3.29$) and the unreliable input group (7 males, 18 females, $M_{age} = 23.12$, $SD = 3.14$).

Participants reported having knowledge of one or more foreign languages. Given that knowledge of foreign languages could facilitate learning in our study, we wanted to check if there were differences between the two experimental groups in this respect. In a questionnaire, learners were asked to list the languages they spoke and assess their level of knowledge (A1 – C2 on the CEFR scales (Council of Europe, 2011)). Based on learners’ reports, we computed two measures of language knowledge, which were created by awarding 0.5 points for each level of each language: 1) the foreign language knowledge (FLK) measure – reflecting the number and level of knowledge of all foreign languages, and 2) the AgrFLK measure – reflecting the number and level of foreign languages that featured gender-based determiners and agreement between determiner and noun. On average, learners scored 9.26 ($SD = 3.58$) on the FLK measure and 2.38 ($SD = 2.17$) on the AgrFLK measure.

A Welch two-sample t-test showed that reliable ($M = 9.28$, $SD = 3.38$) and unreliable input learners ($M = 9.24$, $SD = 3.83$) did not differ significantly on the FLK measure ($t(47.28) = 0.04$, $p = .97$, 95% CI [-2.02, 2.1]). For the AgrFLK measure, a non-parametric Mann-Whitney U test\(^2\) showed that reliable ($M = 2.54$, $SD = 2.2$) and unreliable input learners ($M = 2.22$, $SD = 2.17$) also did not differ significantly ($W = 344$, $p = .54$, 95% CI [-1, 2]).

The experimental session lasted about 1 h and 45 min, and learners received 15 euros for participation. The experiment was approved by the Ethics Committee of the University of Amsterdam.

3.2.2 The target language and exposure
Given that one of our main goals was to replicate findings from the initial study, we used its materials for the present study. Learners were exposed to a miniature language based on Fijian, and the target structure was determiner-noun agreement. Learners first learned the nouns of the target language, after

---

\(^2\)Whenever t-test assumptions (i.e., normality of distribution and homogeneity of variance) were violated, we used a non-parametric Mann-Whitney U test instead of a t-test.
which their knowledge of nouns was tested using a multiple choice noun test. Learners were then exposed to more complex linguistic input featuring the target agreement pattern. For a detailed description of the target language, the determiner-noun agreement pattern, and the language exposure materials, see Chapter 2. In terms of design and materials used, there was only one difference between the initial study and the present study. Learners in the present study were exposed to an additional set of 176 trials in the target structure exposure phase, which means that their exposure to the target pattern was about 60% longer, both in terms of number of trials and duration. This additional set of trials was created by selecting and reordering the first 176 trials of the original set. In total, learners were exposed to 484 trials (lasting 40 minutes) as compared to 308 trials (lasting 25 minutes) in the initial study.

3.2.3 Oral GJT's
After the exposure, we administered two GJT's to test learners' knowledge of the target structure. These tests were identical to the ones administered in the initial study.

3.2.4 Debriefing
After the language exposure and tests, learners were debriefed using the same protocol that we used in the initial, short exposure study. The goal was to find out if learners developed awareness of the target pattern, and if so, at which point in the experiment. We later used the information obtained during debriefing to classify learners as aware or unaware of the target pattern in different stages of the experiment.

3.2.5 Aptitude measures
Participants did three aptitude tests: a non-verbal IQ test, tapping into their analytical ability (Wechsler, 2008) and two working memory tests: the LLAMA D sound recognition task (Meara, 2005) and the visual forward and backward digit span test, which Olsthoorn, Andringa, and Hulstijn (2014) developed based on WAIS-III digit span task (Wechsler, 1997). The IQ and the LLAMA D tests were also used in the initial study; for more details on those tests, see Chapter 2. As in the initial study, we used the noun test scores as a measure of learners' rote memory abilities. Below we present the aptitude measures in the order in which they were administered in the experiment.
The noun test
On average, learners needed 43.38 items ($SD = 5.38$, $min = 40$, $max = 71$) to reach criterion on the noun test, and lower scores indicated quicker learning. We used this test as a measure of learners’ rote memory abilities. However, since the minimum score on this test (i.e., 40) was fixed, the test may not have measured the full range of differences in rote memory between learners. There were 15 learners who scored 40. According to the Mann-Whitney U test, reliable ($M = 44.24$, $SD = 6.67$) and unreliable input learners ($M = 42.52$, $SD = 3.62$) did not differ significantly in noun learning ($W = 381.5$, $p = .18$, 95% CI [-0.0001, 2]).

The non-verbal IQ test
As in the initial study, the test was administered during the first break in the exposure – after 15 minutes of exposure to the target structure. Participants’ average accuracy was 23.14 ($SD = 1.63$, $min = 19$, $max = 26$). A Welch two-sample t-test showed that there were no differences between reliable ($M = 22.96$, $SD = 1.74$) and unreliable input learners ($M = 23.32$, $SD = 1.52$) in their performance on the IQ test ($t(47.12) = -0.78$, $p = .44$, 95% CI [-1.29, 0.57]).

The visual forward and backward digit span task
The digit span task was administered during the second break in the exposure. This was before the last set of exposure trials, i.e. after about 25 minutes of exposure to the target structure. During the digit span task, participants saw sequences of numbers of varying length presented on the screen, and after each sequence they were asked to type in the numbers in the correct order. In the forward digit span task (FDS), participants needed to enter the numbers in the order in which they saw them appear, whereas in the backward digit span task (BDS), they were asked to enter the numbers in reverse order. The sequences were between 2 and 9 digits long in the FDS, and between 2 and 8 digits long in the BDS. Two sequences of each length were presented, and the task stopped automatically once learners made mistakes on two consecutive sequences of the same length. Their score on each task was the length of the longest sequence they could accurately reproduce. On average, learners scored 6.06 on the FDS ($SD = 1.25$, $min = 3$, $max = 9$). Reliable ($M = 6.2$, $SD = 1.32$) and unreliable input learners ($M = 5.92$, $SD = 1.19$) did not differ significantly, as shown by a Mann-Whitney U test ($W = 355.5$, $p = .39$, 95% CI [-0.00003, 1]). On the BDS, learners scored 5.24 ($SD = 1.3$, $min = 3$, $max = 8$). Again, a Welch two-sample t-test showed that reliable input learners ($M = 5.2$, $SD = 1.19$) and unreliable input learners ($M = 5.28$, $SD = 1.43$) did not differ significantly ($t(46.47) = -0.21$, $p = .83$, 95% CI [-0.83, 0.67]). Since the forward and backward digit span scores correlated highly ($r = .63$, $p < .001$), we merged the two tasks.
into a single measure that was created by adding up the BDS and FDS scores and dividing that score by 2. We labeled that measure as *digit span* (DS).

**The LLAMA D sound recognition task**

Participants’ average accuracy on the LLAMA D task was 32.9 (SD = 12.9, min = 0, max = 65). According to a Welch two-sample t-test, there were no differences between the reliable (M = 30.4, SD = 11.89) and unreliable input learners (M = 35.4, SD = 13.61) in their performance on LLAMA D (t(47.15) = -1.38, p = .17, 95% CI [-12.27, 2.27]).

### 3.2.6 Statistical procedures

In this study, we analyzed the results using R (R Development Core Team, 2011) using the same statistical procedures as in the initial study. We centered all continuous variables, such as aptitude scores, so that learners’ individual scores reflected the deviation from all participants’ mean scores. For all categorical variables, we specified explicit contrasts, by assigning -0.5 to one level of the variable and +0.5 to the other level of the variable.

As in the initial study, we used generalized linear models (GLMs) when investigating the relation between learners’ aptitudes and their awareness of the target pattern. For the analysis of GJT performance, we used generalized linear mixed models (GLMMs), which include random effects in addition to fixed effects. The dependent variable was learners’ accuracy, a binomial variable coded as 1 (correct) or 0 (incorrect). Whenever possible, we used maximal models, in which we included all fixed effects under investigation, and all two-way and three-way interactions between them. If a model with a three-way interaction did not converge, we went back to a simpler model with main effects and two-way interactions only.

In our analyses, we followed the recommendation of Barr et al. (2013), and whenever possible, we applied maximal random effects structures. This means that we included both by-subject and by-item random intercepts and slopes for each fixed effect included in the model. When subjects or items were nested within a particular fixed effect variable, we did not include the corresponding by-subject or by-item random slopes because such models would be unidentifiable. If a model with the maximal random structure could not converge, we went back to a simpler model with only by-subject and by-item random intercepts.
3.3 Results – replication of the initial study

In this section, we will analyze the results of the present study in the same way as we did in the initial study in order to see if we can replicate the effects from the initial study. We will start by presenting results related to learners’ pattern awareness. After this, we will present results of the GJT with familiar nouns, where learners were exposed to the nouns they had heard in the exposure. Then we will look at the GJT with novel nouns in order to see if learners could generalize the pattern to the nouns they had not heard in the exposure. Finally, we focus on the relationships between the different aptitude measures and the learning outcomes.

3.3.1 Debriefing results – awareness

Learners reported that they mostly focused on learning nouns, verbs, and adjectives during the exposure. Also, 37 out of 50 learners reported being aware that *lep* was a masculine determiner, whereas *ris* was a feminine determiner. Out of 50 learners, 31 became aware of the agreement pattern during the experiment or immediately after it during debriefing, and could reproduce it correctly. Using the information from the debriefing, we classified learners as aware or unaware of the target pattern at different stages of the experiment. Table 3.1 summarizes the results. As the experiment progressed, more and more learners were becoming aware of the pattern.

<table>
<thead>
<tr>
<th>Stage in the experiment</th>
<th>Reliable input group</th>
<th>Unreliable input group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td>6/25</td>
<td>10/25</td>
</tr>
<tr>
<td>GJT novel nouns</td>
<td>8/25</td>
<td>12/25</td>
</tr>
<tr>
<td>GJT familiar nouns</td>
<td>8/25</td>
<td>12/25</td>
</tr>
<tr>
<td>After experiment</td>
<td>14/25</td>
<td>17/25</td>
</tr>
</tbody>
</table>

3.3.2 GJT with familiar nouns

We analyzed the GJT with familiar nouns to find out if learners acquired the target pattern of the items they had heard in the exposure, and if their knowledge was influenced by the input type (reliable vs. unreliable) they had received. Consistent and inconsistent (i.e., exception) items were analyzed
Effects of prolonged exposure on L2 pattern learning

separately because only consistent items were scored in the same way for the two groups of learners. Inconsistent items were scored differently for the two groups, and accuracy on these items reflected the extent to which learners performed according to what they had heard in the exposure. During this test, 20 learners were aware or became aware of the target pattern (8 in the reliable input group, 12 in the unreliable input group), whereas the others did not report pattern awareness (n = 30; see Table 3.1).

Consistent items analysis
As in the initial study, we included only non-biological gender nouns in the analysis (n = 24) given that learners were more accurate on biological (M = 79.00, SD = 22.51) than on non-biological gender nouns (M = 70.08, SD = 21.78), as shown by a Mann-Whitney U test (W = 1594, p = .01, 95% CI [0.00002, 0.12]). The reason for this was that biological gender nouns (e.g., man, woman, boy, girl) inherently point to gender differences, so learners could perform accurately on them without having knowledge of the agreement pattern or without having memorized determiner-noun combinations.

Average accuracy on non-biological consistent items was 70.08% (SD = 21.78). Overall, learners performed significantly above chance (b = 2.14, SE = 0.58, p < .001, 95% CI [1, 3.38], OR = 8.47). This overall learning effect, with an odds ratio (OR) of 8.47, was larger compared to the initial short exposure study, where the odds ratio was 3.53. In order to find out how unreliable input affected learners’ performance, and whether this differed for aware and unaware learners, we built a model that included the following fixed effects: group (reliable vs. unreliable input group), pattern awareness (whether or not learners were aware of the target pattern during the test), and grammaticality of items (grammatical vs. ungrammatical). In this model, we included only by-subject and by-item random intercepts. Random slopes for fixed effects could not be included due to non-convergence.

We found no main effect of group (b = 0.46, SE = 0.52, p = .37, 95% CI [-0.57, 1.49], OR = 1.59). There was a significant effect of awareness (b = 2.53, SE = 0.52, p < .001, 95% CI [1.5, 3.56], OR = 12.57) – learners who were aware of the pattern were more accurate than learners who did not report such awareness. We also found an effect of grammaticality (b = -3.05, SE = 0.37, p < .001, 95% CI [-3.79, -2.3], OR = 0.05) – learners were less accurate on ungrammatical items, which implies that they had a bias for marking items as grammatical. When it comes to the two-way interactions, the following were significant: group x grammaticality (b = -1.18, SE = 0.5, p = .02, 95% CI [-2.18, -0.17], OR = 0.31) – learners in the unreliable group were less accurate on ungrammatical items than learners in the reliable input group, and awareness x grammaticality (b = 1.88, SE = 0.5, p < .001, 95% CI [0.88, 2.89], OR = 6.58) –
learners aware of the pattern performed better on ungrammatical items than learners without pattern awareness. The interaction between group and awareness did not reach significance although aware learners tended to be less accurate if they had received unreliable compared to reliable input (Table 3.2; Figure 3.1) \( (b = -1.73, SE = 1.03, p = .09, 95\% \text{ CI } [-3.79, 0.34], OR = 0.18) \). Importantly, there was a three-way interaction between group, awareness, and grammaticality (Table 3.2), meaning that learners were less accurate on ungrammatical items when they were aware of the pattern and had received unreliable input \( (b = -3.59, SE = 1, p < .001, 95\% \text{ CI } [-5.59, -1.58], OR = 0.03) \).

Next, we looked at whether the negative effects of unreliable input could also be observed from aware learners’ sources of reliance during the test, as reported during debriefing. We found that this was the case: in the reliable input group, out of 8 learners aware of the pattern, 7 reportedly relied on it during the test. However, in the unreliable input group, out of 12 learners aware of the pattern, only 4 reportedly relied on it.

If we compare the results presented above with the results of the initial study, we can see that findings of the initial study that pertain to learners aware of the pattern have been replicated. When learners were aware of the pattern, unreliable input disturbed their performance. The only difference is that in the present study, this effect was especially visible on learners’ performance on ungrammatical items, which can be seen from the three-way interaction between group, grammaticality, and awareness. Similar to the initial study, learners differed substantially in whether and to what extent they were disturbed by exceptions in the input (Figure 3.1; SDs in Table 3.2). Also, as debriefing results showed, learners who had received unreliable input often did not feel sufficiently confident to rely on the pattern they had noticed.

**Table 3.2.** Overview of learners’ mean accuracies and their standard deviations (SD), split by group, pattern awareness, and grammaticality on the GJT with familiar nouns – consistent items

<table>
<thead>
<tr>
<th>Grammatical items</th>
<th>Ungrammatical items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aware learners</td>
</tr>
<tr>
<td><strong>Group</strong></td>
<td>% correct (SD)</td>
</tr>
<tr>
<td><strong>Reliable</strong></td>
<td>94.79 (7.63)</td>
</tr>
<tr>
<td><strong>Unreliable</strong></td>
<td>96.53 (5.57)</td>
</tr>
</tbody>
</table>
Figure 3.1. Overview of accuracies on the GJT familiar nouns – consistent items, for each individual learner, split by group and pattern awareness

Next, we wanted to find out if learners unaware of the pattern performed above chance, and if their performance was affected by unreliable input. For this, we ran a separate analysis on unaware learners only. These learners obtained an average score of 58.75% ($SD = 16.28$). There was a trend towards above chance performance ($b = 0.9, SE = 0.57, p = .11, 95\% CI [-0.23, 2.11], OR = 2.46$). We built a model with group and grammaticality as fixed effects. There was a main effect of group, such that unreliable input learners were overall more accurate than reliable input learners (Table 3.2; Figure 3.1) ($b = 2.01, SE = 0.74, p = .006, 95\% CI [0.53, 3.49], OR = 7.47$), and a main effect of grammaticality – learners were less accurate on ungrammatical items ($b = -5.98, SE = 1.24, p < .001, 95\% CI [-8.45, -3.5], OR = 0.003$). We found no interaction between group and grammaticality ($b = 2.7, SE = 2.04, p = .19, 95\% CI [-1.38, 6.77], OR = 14.86$).

In order to identify individual unaware learners who performed reliably above chance, we used a binomial test, which tests if a score is above chance given the number of trials. Learners who scored 75% or above could be classified as having performed above chance ($p < .05$). We found 6 learners who were unaware of the pattern during the test but performed reliably above chance. Interestingly, all these learners were in the unreliable input group. Five of them reported having relied on memory, and 1 reported reliance of intuition.

In the initial short exposure study, when learners were unaware of the pattern, we found no evidence that unreliable input had any effect on their performance. In the present study, we observed a different result: unaware learners who had heard exceptions in the input outperformed learners who had
heard fully regular input, which may imply that unreliable input with exceptions fostered better memorizing of consistent items for unaware learners. It should be noted that there were considerable differences between unaware learners in whether and to what extent they benefited from unreliable input (Figure 3.1; SDs in Table 3.2).

Inconsistent items analysis
Next, we analyzed learners’ performance on inconsistent items, which were exceptions only for the unreliable input group. The goal was to check if learners’ performance on the GJT with familiar nouns was based on memorization of the determiner-noun combinations from the input, or on more abstract pattern knowledge. Accuracy on the inconsistent items reflected what participants in each group had heard during the exposure. Just as in the initial study, we assumed that if the unreliable input learners were less accurate on inconsistent items than the reliable input learners, this would be evidence of overgeneralizing the pattern to exceptions and pattern learning, rather than just memorizing determiner-noun combinations (Tomasello, 2000).

Accuracy of the reliable input learners was 69% (SD = 21.98) on the inconsistent items, whereas the average accuracy of the unreliable input learners was 57.5% (SD = 28.18). Overall, learners performed significantly above chance \( (b = 0.68, SE = 0.19, p < .001, 95\% CI [0.3, 1.06], OR = 1.98) \). We then ran a model including group, awareness, and grammaticality as fixed effects. Given that the model with the three-way interaction did not converge, we had to exclude that interaction from the model. We found a main effect of group, with unreliable input learners being overall less accurate than reliable input learners \( (b = -3.8, SE = 1.09, p < .001, 95\% CI [-5.97, -1.62], OR = 0.02) \). There were also main effects of pattern awareness – aware learners were more accurate than unaware learners \( (b = 3.53, SE = 1.01, p < .001, 95\% CI [1.5, 5.54], OR = 34) \), and grammaticality \( (b = -4.79, SE = 0.79, p < .001, 95\% CI [-6.36, -3.22], OR = 0.01) \) – learners were less accurate on ungrammatical items. The following two-way interactions were significant: group x awareness – when aware learners had heard unreliable input they were less accurate than when they had heard reliable input (Table 3.3; Figure 3.2) \( (b = -8.58, SE = 2.3, p < .001, 95\% CI [-13.19, -3.98], OR = 0.0002) \), and group x grammaticality – unreliable input learners were more accurate on ungrammatical items than reliable input learners \( (b = 4.1, SE = 1.52, p = .007, 95\% CI [1.05, 7.14], OR = 60.06) \). The interaction between awareness and grammaticality was not significant \( (b = -0.59, SE = 1.44, p = .68, 95\% CI [-3.46, 2.28], OR = 0.56) \).

The results presented above show that aware unreliable input learners were less accurate on inconsistent items than aware reliable input learners. This suggests that aware unreliable input learners did not just follow what they
had heard in the input, but overgeneralized the pattern they had noticed to exceptions, which confirms that they learned the underlying pattern. This result fully replicates the result observed in the initial short exposure study.

Table 3.3. Overview of learners’ mean accuracies and their standard deviations (SD), split by group, pattern awareness, and grammaticality on the GJT with familiar nouns – inconsistent items

<table>
<thead>
<tr>
<th></th>
<th>Grammatical items</th>
<th>Ungrammatical items</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aware learners</td>
<td>Unaware learners</td>
<td>Aware learners</td>
<td>Unaware learners</td>
<td>Aware learners</td>
<td>Unaware learners</td>
</tr>
<tr>
<td>Group</td>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
</tr>
<tr>
<td>Reliable</td>
<td>100 (0)</td>
<td>94.12 (10.93)</td>
<td>90.63 (18.6)</td>
<td>19.12 (30.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unreliable</td>
<td>72.92 (36.08)</td>
<td>80.77 (20.8)</td>
<td>33.33 (43.08)</td>
<td>42.31 (38.71)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.2. Overview of accuracies on the GJT familiar nouns – inconsistent items, for each individual learner, split by group and pattern awareness

Next, we wanted to check if learners who were unaware of the pattern showed any signs of overgeneralization of the pattern (i.e., implicit pattern learning) as we had observed in the initial short exposure study. We ran an analysis on the data of the unaware learners only, and we included group and grammaticality as fixed effects in the model. We found no main effect of group ($b = 0.34, SE = 0.56, p = .54, 95\% CI [-0.78, 1.47], OR = 1.41$). There was a
significant main effect of grammaticality \((b = -4.02, SE = 0.74, p < .001, 95\% CI [-5.49, -2.55], OR = 0.02)\) – learners were less accurate on ungrammatical items, and a significant interaction between group and grammaticality, such that the unreliable input learners were more accurate on ungrammatical items than the reliable input learners (Table 3.3 above) \((b = 3.52, SE = 1.47, p = .017, 95\% CI [0.58, 6.46], OR = 33.87)\).

If we compare the results pertaining to unaware learners in the present study with the results we found in the initial study, we observe clear differences. In the initial study, we found evidence of implicit pattern learning because learners in the unreliable group overgeneralized exceptions, and were less accurate than learners in the reliable group. Surprisingly, in this study, unreliable input learners did not have lower accuracy scores due to overgeneralization, but instead they had higher accuracy scores than reliable input learners when responding to inconsistent items. This means not only that there was no evidence of implicit pattern learning among unaware learners in this study, but it also suggests that unreliable input fostered better memorizing of the input. This effect is especially visible from the better performance of unreliable input learners on ungrammatical items, compared to the reliable input learners (Table 3.3). Also, note that in the consistent items analysis presented above, we similarly observed that unreliable input led to better memorizing of items among unaware learners.

### 3.3.3 GJT with novel nouns

The goal of this task was to find out if learners could generalize the target pattern to novel nouns that they had not heard during exposure, and to see if this ability was affected by the kind of input they had been exposed to. In the initial study, we found that only learners aware of the pattern performed above chance on this task, and that unreliable input disturbed these learners’ performance. We wanted to see if we could replicate this result in the present study. During this task, 20 learners were or became aware of the target pattern. Eight were in the reliable input group, and 12 were in the unreliable input group.

The average accuracy on the test was 61.7\% \((SD = 26.89)\). Learners as a group performed significantly above chance \((b = 0.74, SE = 0.23, p = .001, 95\% CI [0.3, 1.2], OR = 2.1)\), and this learning effect was similar to the one in the initial study \((OR = 1.97)\). We were interested in whether unreliable input affected learners’ ability to generalize the pattern to novel nouns, and whether this differed for aware and unaware learners. Therefore, we built a model with the following fixed effects: group, grammaticality, and awareness of the pattern. We found main effects of group \((b = -0.62, SE = 0.24, p = .001, 95\% CI [-1.1, 0.14], OR = 0.54)\) – learners in the unreliable input group were overall less
Effects of prolonged exposure on L2 pattern learning

accurate than learners in the reliable input group; grammaticality \((b = -1.31, SE = 0.27, p < .001, 95\% \text{ CI } [-1.84, -0.77], OR = 0.27)\) – learners were less accurate on ungrammatical items; and awareness \((b = 2.07, SE = 0.24, p < .001, 95\% \text{ CI } [1.59, 2.54], OR = 7.89)\) – learners were more accurate if they were aware of the target pattern. We found a significant interaction between group and awareness \((b = -1.16, SE = 0.48, p = .015, 95\% \text{ CI } [-2.11, -0.2], OR = 0.31)\), such that aware learners were less accurate if they had heard unreliable input (Table 3.4; Figure 3.3). The other two-way interactions were not significant: group x grammaticality \((b = -0.19, SE = 0.44, p = .67, 95\% \text{ CI } [-1.06, 0.69], OR = 0.83)\) and awareness x grammaticality \((b = 0.31, SE = 0.43, p = .47, 95\% \text{ CI } [-0.55, 1.18], OR = 1.37)\). The three-way interaction group x grammaticality x awareness was also not significant \((b = -1.27, SE = 0.87, p = .14, 95\% \text{ CI } [-3.01, 0.49], OR = 0.28)\) although there was a slight tendency that learners were less accurate on ungrammatical items when they were aware of the pattern and had received unreliable input (Table 3.4).

Learners’ reports during debriefing confirmed that learners who had heard unreliable input often did not feel confident to rely on the pattern they had noticed. In the reliable input group, out of 8 learners aware of the pattern, 7 reported having relied on it during the test, whereas in the unreliable input group, out of 12 learners aware of the pattern, only 6 relied on it.

The results presented above fully replicated the results of the initial study: aware learners were able to generalize the pattern they had noticed to novel nouns, and their ability to do so was negatively affected by the presence of exceptions in the input. Again, we observed individual differences in that some aware learners could cope with unreliable input better than the others (Figure 3.3; SDs in Table 3.4).

Table 3.4. Overview of learners’ mean accuracies and their standard deviations (SD), split by group, pattern awareness, and grammaticality on the GJT with novel nouns

<table>
<thead>
<tr>
<th></th>
<th>Grammatical items</th>
<th>Ungrammatical items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aware learners</td>
<td>Unaware learners</td>
</tr>
<tr>
<td><strong>Group</strong></td>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
</tr>
<tr>
<td>Reliable</td>
<td>92.5 (13.63)</td>
<td>67.65 (12.88)</td>
</tr>
<tr>
<td>Unreliable</td>
<td>86.25 (14.79)</td>
<td>61.54 (17.6)</td>
</tr>
</tbody>
</table>
The analysis above could not tell us if learners unaware of the pattern performed above chance, and whether their performance was disturbed by unreliable input. We therefore ran a separate analysis on the results of unaware learners only (\( M = 49.33, SD = 21.58 \)). We found no above chance performance (\( b = -0.03, SE = 0.14, p = 0.85, 95\% CI [-0.31, 0.25], OR = 0.97 \)). There was no effect of group (\( b = -0.06, SE = 0.13, p = .66, 95\% CI [-0.32, 0.2], OR = 0.94 \)), but there was a significant main effect of grammaticality (\( b = -1.41, SE = 0.28, p < .001, 95\% CI [-1.97, -0.86], OR = 0.24 \)). The interaction between group and grammatically was not significant (\( b = 0.42, SE = 0.47, p = .38, 95\% CI [-1.36, 0.53], OR = 1.52 \)).

Finally, we checked if there were any learners unaware of the pattern who performed significantly above chance on the test. According to the binomial test, learners who scored 67.5\% or more could be reliably classified as having performed above chance (\( p < 0.05 \)), but we found no such learners in this study.

These results replicate the initial study results, where we also found no overall evidence of unaware learners being able to generalize pattern to novel items. The only difference is that in the initial study, there were two learners who were unaware of the pattern but showed above chance performance on the GJT with novel nouns.

### 3.3.4 The role of working memory, rote memory, and analytical ability

Next, we were interested in whether the working memory, rote memory, and analytical ability measures predicted learners’ performance on the two GJT's,
Effects of prolonged exposure on L2 pattern learning

and whether this differed for the two instruction groups. We also wanted to check if the aptitude measures could predict whether or not learners would become aware of the pattern.

First, we looked at the correlations between the different aptitude measures, and we found significant correlations between LLAMA D and noun test scores \((r = -0.36, p = .01)\), and between the digit span task (DS measure) and the noun test scores \((r = -0.44, p = .001)\). Also, there were significant correlations between IQ scores and the digit span task \((r = 0.35, p = .01)\), as well as between IQ and the noun test scores \((r = -0.28, p = .046)\).

All tests that tap into learners’ memory capacities predicted overall performance on the GJT with familiar nouns – consistent, non-biological items (Table 3.5; Figure 3.4): LLAMA D, digit span, and noun test scores. This means that if learners had higher scores on LLAMA D and digit span task, and needed fewer items to pass the noun test, they were also more accurate on the GJT with familiar nouns. Learners’ analytical abilities (IQ measure) did not predict overall performance on the GJT with familiar nouns. We found no aptitude by treatment (reliable vs. unreliable input) interactions (Table 3.5).
Table 3.5. Overview of aptitude involvement in learners’ performance on the GJT with familiar nouns, GJT with novel nouns, and learners’ pattern awareness

<table>
<thead>
<tr>
<th></th>
<th>GJT WITH FAMILIAR NOUNS</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( b )</td>
<td>( SE )</td>
<td>( p )</td>
<td>( 95% CI )</td>
<td>( OR )</td>
</tr>
<tr>
<td><strong>Main effects of aptitude measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D</td>
<td>0.07</td>
<td>0.02</td>
<td>&lt; .001</td>
<td>[0.04, 0.11]</td>
<td>1.08</td>
</tr>
<tr>
<td>Noun test</td>
<td>-0.11</td>
<td>0.03</td>
<td>.001</td>
<td>[-0.18, -0.04]</td>
<td>0.9</td>
</tr>
<tr>
<td>Digit span</td>
<td>0.86</td>
<td>0.24</td>
<td>&lt; .001</td>
<td>[0.39, 1.34]</td>
<td>2.37</td>
</tr>
<tr>
<td>IQ</td>
<td>0.3</td>
<td>0.2</td>
<td>.14</td>
<td>[-0.1, 0.7]</td>
<td>1.35</td>
</tr>
<tr>
<td><strong>Aptitude ( \times ) group interactions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D ( \times ) group</td>
<td>0.04</td>
<td>0.03</td>
<td>.23</td>
<td>[-0.03, 0.11]</td>
<td>1.04</td>
</tr>
<tr>
<td>Noun test ( \times ) group</td>
<td>-0.1</td>
<td>0.1</td>
<td>.3</td>
<td>[-0.31, 0.1]</td>
<td>0.9</td>
</tr>
<tr>
<td>Digit span ( \times ) group</td>
<td>0.02</td>
<td>0.45</td>
<td>.97</td>
<td>[-0.89, 0.92]</td>
<td>1.02</td>
</tr>
<tr>
<td>IQ ( \times ) group</td>
<td>-0.4</td>
<td>0.35</td>
<td>.24</td>
<td>[-1.09, 0.29]</td>
<td>0.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GJT WITH NOVEL NOUNS</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( b )</td>
<td>( SE )</td>
<td>( p )</td>
<td>( 95% CI )</td>
<td>( OR )</td>
</tr>
<tr>
<td><strong>Main effects of aptitude measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D</td>
<td>0.02</td>
<td>0.01</td>
<td>.04</td>
<td>[0.0004, 0.04]</td>
<td>1.02</td>
</tr>
<tr>
<td>Noun test</td>
<td>-0.05</td>
<td>0.02</td>
<td>&lt; .001</td>
<td>[-0.08, -0.02]</td>
<td>0.95</td>
</tr>
<tr>
<td>Digit span</td>
<td>0.22</td>
<td>0.16</td>
<td>.16</td>
<td>[-0.09, 0.53]</td>
<td>1.25</td>
</tr>
<tr>
<td>IQ</td>
<td>0.19</td>
<td>0.17</td>
<td>.26</td>
<td>[-0.15, 0.53]</td>
<td>1.21</td>
</tr>
<tr>
<td><strong>Aptitude ( \times ) group interactions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D ( \times ) group</td>
<td>0.004</td>
<td>0.02</td>
<td>.82</td>
<td>[-0.03, 0.04]</td>
<td>1.004</td>
</tr>
<tr>
<td>Noun test ( \times ) group</td>
<td>0.05</td>
<td>0.06</td>
<td>.43</td>
<td>[-0.07, 0.16]</td>
<td>1.05</td>
</tr>
<tr>
<td>Digit span ( \times ) group</td>
<td>-0.65</td>
<td>0.31</td>
<td>.03</td>
<td>[-1.27, -0.04]</td>
<td>0.52</td>
</tr>
<tr>
<td>IQ ( \times ) group</td>
<td>-0.48</td>
<td>0.2</td>
<td>.015</td>
<td>[-0.88, -0.08]</td>
<td>0.62</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PATTERN AWARENESS</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( b )</td>
<td>( SE )</td>
<td>( p )</td>
<td>( 95% CI )</td>
<td>( OR )</td>
</tr>
<tr>
<td><strong>Main effects of aptitude measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D</td>
<td>0.09</td>
<td>0.03</td>
<td>.004</td>
<td>[0.04, 0.16]</td>
<td>1.1</td>
</tr>
<tr>
<td>Noun test</td>
<td>-0.16</td>
<td>0.08</td>
<td>.06</td>
<td>[-0.35, -0.02]</td>
<td>0.85</td>
</tr>
<tr>
<td>Digit span</td>
<td>0.09</td>
<td>0.26</td>
<td>.73</td>
<td>[-0.42, 0.6]</td>
<td>1.09</td>
</tr>
<tr>
<td>IQ</td>
<td>0.18</td>
<td>0.18</td>
<td>.31</td>
<td>[-0.17, 0.56]</td>
<td>1.2</td>
</tr>
</tbody>
</table>
Figure 3.4. Relation between learners’ predicted performance on the GJT with familiar nouns and a) the LLAMA D scores, b) the noun test scores, and c) the digit span scores.

Learners’ overall accuracy on the GJT with novel nouns was predicted by learners’ working memory (LLAMA D) and rote memory (noun test) (Table 3.5; Figure 3.5). We also found significant aptitude by treatment interactions. There were interactions between digit span scores and group, as well as between IQ scores and group (Table 3.5). Judging from Figure 3.6, these interactions mean that higher analytical ability and working memory were associated with higher GJT accuracy for reliable input learners only.
Finally, we were interested in whether any of the aptitudes measured could predict if learners would develop awareness of the pattern. We found that LLAMA D and noun test scores predicted awareness in that learners with higher LLAMA D scores and better noun test performance were more likely to become aware of the pattern (Table 3.5).

If we compare the presented results on the role of aptitudes in learning with the results obtained in the initial short exposure study, we can see that the present study partially replicated previously obtained results. Generally speaking, the present study confirms the involvement of both analytical ability and working memory in learning, which was also observed in the initial study. However, the picture we found in this study somewhat differs from the results of the initial study. Table 3.6 summarizes the similarities and the differences.
Effects of prolonged exposure on L2 pattern learning

We found that in both the present and the initial study, learners’ LLAMA D and noun scores predicted their performance on the GJT with familiar nouns. However, in the initial study, also links with IQ scores were observed, which is not replicated in the present study.

As for learners’ ability to generalize the pattern to novel items, in both the present and the initial study, we found that higher LLAMA D scores predicted higher scores on the GJT with novel nouns. However, while this link pertained to learners as a group in the present study, it was restricted to reliable input learners in the initial study. Also, in the present study, we observed interactions between IQ and digit span on the one hand, and group on the other hand, such that higher IQ and digit span scores were associated with better performance on novel GJT for reliable input learners only.

Finally, when it comes to the question of which aptitudes could predict awareness, this study replicated the result that LLAMA D scores predicted learners’ pattern awareness. However, unlike in the initial study, we found no links between pattern awareness and IQ scores.

Table 3.6. Overview of the involvement of memory measures (M) and analytical ability measures (AA) in the familiar and novel GJT performance and pattern awareness for the initial study (Chapter 2) and the present study

<table>
<thead>
<tr>
<th></th>
<th>GJT familiar nouns</th>
<th>GJT novel nouns</th>
<th>Pattern awareness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>AA</td>
<td>M</td>
</tr>
<tr>
<td>Initial study</td>
<td>LLAMA D</td>
<td>IQ</td>
<td>LLAMA D x group</td>
</tr>
<tr>
<td></td>
<td>Noun test</td>
<td>Noun test</td>
<td></td>
</tr>
<tr>
<td>This study</td>
<td>LLAMA D</td>
<td>none</td>
<td>LLAMA D</td>
</tr>
<tr>
<td></td>
<td>Noun test</td>
<td>Digit span</td>
<td>Noun test</td>
</tr>
</tbody>
</table>

3.4 Results – effects of longer exposure on implicit learning and awareness

Our ultimate goal was to statistically compare the results from the initial short exposure study and the present study in order to see whether prolonged exposure led to more awareness of the target pattern or to better implicit learning, or both. Given that we wanted to compare two data samples that were collected at different moments in time, it was important to first compare the short and long exposure learners on different background and aptitude
measures in order to make sure that the differences observed came from differences in the length of exposure rather than other variables.

First, we checked if the two groups of learners differed on the FLK and AgrFLK measures of knowledge of foreign languages. A Welch two-sample t-test showed that there were no differences between short exposure (M = 8.15, SD = 3.49) and long exposure learners (M = 9.26, SD = 3.58) in their knowledge of foreign languages – FLK measure (t(97.95) = 1.57, p = .12, 95% CI [-0.29, 2.51]). Also, learners in the short exposure group (M = 1.91, SD = 2.44) and long exposure group (M = 2.38, SD = 2.17) did not differ in their knowledge of languages with agreement (AgrFLK measure), as shown by a Mann-Whitney U test (W = 1481.5, p = .10, 95% CI [-0.00002, 1]).

When it comes to the different aptitude measures, learners in the short (M = 22.84, SD = 2.06) and long exposure groups (M = 23.14, SD = 1.63) did not differ in their IQ test performance (W = 1358.5, p = .45, 95% CI [-0.00003, 1]). However, we found that learners in the long exposure group performed significantly better on the noun test and the LLAMA D test, which may indicate that the long exposure learners had better memory capacities. A Mann-Whitney U test showed that long exposure learners (M = 43.38, SD = 5.38) were able to pass the noun test significantly faster than short exposure learners (M = 45.72, SD = 6.32, W = 845.5, p = .005, 95% CI [-3, -0.0001]). Similarly, long exposure learners (M = 32.9, SD = 12.9) performed better than short exposure learners (M = 22.8, SD = 12.78) on the LLAMA D test (t(97.99) = 3.93, p < .001, 95% CI [5.003, 15.2]). Because of these differences, we added LLAMA D and noun test scores as covariates in the subsequent analyses, as a way of controlling for the observed differences in memory.

First, we looked at whether longer exposure led to more learners becoming aware of the target pattern (Table 3.7). For the analysis, we used a GLM, where the dependent variable was awareness of the pattern, and the fixed effect was the exposure length (long vs. short). First we looked at learners’ overall awareness, regardless of when it occurred (i.e., during the exposure, the GJTs, or the debriefing) given that it is possible that exposure differences resulted in awareness differences that were not only visible during the exposure itself but potentially also later during the tests or debriefing. We found a significant effect of exposure length, such that learners were more likely to become aware of the pattern if they had received long compared to short exposure (b = 0.98, SE = 0.41, p = .017, 95% CI [0.16, 1.8], OR = 2.66). However, once we included LLAMA D and noun test scores as covariates in the model in order to control for learners’ between-group differences on these two aptitude measures, this effect disappeared (b = 0.31, SE = 0.48, p = .51, 95% CI [-0.65, 1.27], OR = 1.36). This implies that the observed awareness differences occurred due to better memory capacities of learners in the long exposure group. We noticed that learners’ working memory capacities (LLAMA D) were mostly driving the
differences in awareness ($b = 0.06, SE = 0.02, p = .002, 95\% CI [0.02, 0.1], OR = 1.07$). Next, we performed the same analysis, but zoomed in on the awareness that learners developed during exposure. We observed very similar results. We first found a tendency that longer exposure led to more learners becoming aware of the pattern ($b = 0.9, SE = 0.49, p = .07, 95\% CI [-0.08, 1.89], OR = 2.47$), but once LLAMA D and noun test scores were added as covariates, this tendency disappeared ($b = 0.38, SE = 0.54, p = .48, 95\% CI [-0.7, 1.46], OR = 1.46$). Only LLAMA D scores predicted whether learners would develop awareness of the pattern ($b = 0.05, SE = 0.02, p = .01, 95\% CI [0.01, 0.1], OR = 1.06$).

Table 3.7. Overview of the cumulative number of learners aware of the target pattern at different stages in the experiment, split by exposure length (short vs. long)

<table>
<thead>
<tr>
<th>Stage in the experiment</th>
<th>Short exposure</th>
<th>Long exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td>8/50</td>
<td>16/50</td>
</tr>
<tr>
<td>GJT novel nouns</td>
<td>13/50</td>
<td>20/50</td>
</tr>
<tr>
<td>GJT familiar nouns</td>
<td>16/50</td>
<td>20/50</td>
</tr>
<tr>
<td>After experiment</td>
<td>19/50</td>
<td>31/50</td>
</tr>
</tbody>
</table>

Finally, we were interested in comparing implicit pattern learning in the short versus long exposure studies in order to find out if learners showed higher levels of implicit pattern learning when they received longer exposure to the target pattern. For this purpose, we looked for evidence of overgeneralization of the pattern on exception items by unaware learners. We built a GLMM with LLAMA D and noun test scores as covariates. The dependent variable was accuracy, and the fixed effects were exposure length (long vs. short), grammaticality, and group (reliable vs. unreliable input). There was a main effect of grammaticality ($b = -2.53, SE = 0.49, p < .001, 95\% CI [-3.52, -1.55], OR = 0.08$), but no main effect of group ($b = -0.34, SE = 0.47, p = .47, 95\% CI [-1.28, 0.6], OR = 0.71$), or exposure length ($b = -0.09, SE = 0.36, p = .8, 95\% CI [-0.81, 0.63], OR = 0.91$). There was a significant interaction between exposure length and grammaticality – learners who had received long exposure were less accurate on ungrammatical items ($b = -2.38, SE = 0.8, p = .003, 95\% CI [-3.98, -0.78], OR = 0.09$). The other two-way interactions were not significant: group x grammaticality ($b = 0.97, SE = 0.79, p = .22, 95\% CI [-0.61, 2.55], OR = 2.64$), and exposure length x group ($b = 0.7, SE = 0.6, p = .24, 95\% CI [-0.49, 1.89], OR =
Importantly, we found a significant three-way interaction between exposure length, group, and grammaticality (Table 3.8) indicating that when learners received longer exposure and unreliable input, they were more accurate at judging ungrammatical items ($b = 4.34$, $SE = 1.69$, $p = .01$, 95% CI [0.95, 7.72], $OR = 76.52$). This result suggests that learners who were exposed to unreliable input more intensively were better at recognizing determiner-noun combinations the way they had heard them in the input, which implies not only that unaware long exposure learners did not overgeneralize exceptions but they even showed stronger memorizing. However, this advantage in memorizing was only present when learners had also heard unreliable input.

### Table 3.8. Overview of unaware learners' mean accuracies and their standard deviations (SD), split by group, exposure length, and grammaticality on the GJT with familiar nouns – inconsistent items

<table>
<thead>
<tr>
<th></th>
<th>Grammatical items</th>
<th>Ungrammatical items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Short exposure</td>
<td>Long exposure</td>
</tr>
<tr>
<td>Group</td>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
</tr>
<tr>
<td>Reliable</td>
<td>79.69 (20.85)</td>
<td>94.12 (10.93)</td>
</tr>
<tr>
<td>Unreliable</td>
<td>73.61 (26.39)</td>
<td>80.77 (20.8)</td>
</tr>
</tbody>
</table>

### 3.5 Discussion

This study had two goals. The main one was to investigate how prolonging of exposure influences L2 pattern learning, i.e. whether it leads to more learners becoming aware of the underlying pattern, to higher levels of implicit pattern learning, or possibly both. The second goal was to try and replicate the results of the initial study reported in Chapter 2. We will first discuss to what extent the present study replicated the findings observed in the initial study, after which we will move to the discussion of our main research question related to the effects of exposure length on L2 pattern learning.

Many researchers in the SLA field have pointed out the importance of conducting replication studies in order to validate research results and help distinguish the spurious from real effects (e.g., Handley & Marsden, 2014; Mackey, 2012; Porte, 2010; Polio & Gass, 1997). This study aimed to check if the results obtained in the Chapter 2 study would be replicated. This is why in the present study we again looked at how unreliable input with exceptions...
affects L2 learning of determiner-noun agreement pattern, and how this learning may be predicted by learners’ working memory and analytical abilities.

For learners aware of the pattern, we found very similar results in the present and the initial study, irrespective of the differences in exposure length. In both studies, only learners aware of the target pattern showed above chance performance on the GJT with novel nouns, and these learners’ performance on both GJTs with novel and familiar nouns was disturbed by the presence of exceptions in the input. Learners’ reports on the sources of reliance during GJTs in both studies indicated that exceptions often led learners to feel less confident and avoid relying on the pattern they had noticed. It is also important to note that learners in both studies differed substantially in whether and to what extent they were disturbed by exceptions in the input. This means that some pattern aware learners were strongly disturbed by exceptions, while others still managed to achieve very high accuracies comparable to the ones found among reliable input learners. The results of our studies underline the need to consider individual differences between learners in language research, rather than presenting only group comparisons (Dabrowska & Street, 2006).

For learners who reported no pattern awareness, we observed substantial differences between the present study and the initial study, and these differences may have resulted from differences in the length of exposure. In the short exposure study, we found evidence of pattern learning without awareness, i.e. unaware learners overgeneralized the pattern to exceptions. However, in the present study, there was no evidence of similar implicit pattern learning. Instead, we observed only effects of memorizing, and we noticed that learners who had heard unreliable input were better at memorizing determiner-noun combinations the way they had heard them in the input. In the short exposure study, we found no evidence that unreliable input had any effect on unaware learners’ GJT performance.

This replication study is itself in need of replication, but if its outcomes hold, the implications are substantial. They mean that exposure length and the moment at which learners’ knowledge is tested during the exposure may lead to substantial differences in what kind of performance is observed. In our data, for learners who were unaware of the underlying patterns, shorter exposure and early testing revealed signs of implicit knowledge, whereas longer exposure and later testing only showed effects of memorization, which potentially take some time and exposure to become visible. This may imply that implicit learning effects can only be found if testing is done early enough, in relatively beginning stages of learning. When learners receive longer exposure and are tested later, pattern learning effects may no longer be visible because they get overridden by memory effects. In these cases, pattern knowledge may still be tested using novel items that learners had not been exposed to.
However, as shown in our study, implicit learning effects may be too subtle to appear in tests with novel items given that such tests may be too difficult, and can make learners feel confused and uncomfortable.

When it comes to aware learners, our results suggest that once learners have developed awareness, they base their behavior on explicit knowledge, and additional input and later testing no longer influence the kind of knowledge and performance observed. The finding that testing moment may matter for what kind of knowledge is measured in unaware learners (i.e., memory or pattern-based) may have important methodological implications for the field of second language acquisition, particularly for studies that aim to find evidence of implicit pattern learning. Our results suggest that such studies need to carefully consider the length of exposure and moment of testing, or even include these as variables in their designs. As indicated, the results we observed are themselves in need of replication with an improved design, where learners are assigned to different lengths of exposure randomly or perhaps in a longitudinal design.

In both the present and the initial study, we found links between learning of the target structure and both working memory and analytical ability. However, the exact results we observed in the present study somewhat differed from the ones observed in the initial study (see Table 3.6). These differences may come from differences in the exposure length, as well as from the effects of later testing. Another possibility is that some of the effects or interactions that were present in one but not in the other study are chance findings. However, broadly speaking, if we take the two studies together, we can see that working memory and analytical ability are involved in learning and developing awareness of determiner-noun agreement, and that this involvement may depend on instructional treatment, i.e. it may be stronger when learners are exposed to reliable compared to unreliable input. However, based on our results, it is still difficult to say whether the relationship between aptitudes and learning was necessarily causal. As already pointed out in the initial study, such evidence can come from studies similar to Wesche (1981), where learners are assigned to particular treatment groups based on their aptitude profiles, and as a result of such assignment, a relationship is observed between aptitudes and learning.

The main goal of this study was to provide some novel insights on how awareness and implicit pattern learning develop from shorter vs. longer exposure to language. The results could also give us the possibility to speculate - although in an indirect way - about the relation between awareness and implicit learning, and how implicit learning may influence awareness, a topic that has received very little attention in the literature. We found that longer exposure to the target pattern led to the disappearance of implicit pattern learning signs and appearance of the effects of memorization. With respect to awareness, we found that more learners developed awareness in the long
exposure study (31 out of 50) compared to the short exposure study (19 out of 50). At first glance, this result may be seen as in line with both Cleeremans (2011) and Bialystok (1989, 1994), who assume that awareness develops as a result of accumulated implicit knowledge (Cleeremans, 2011) or as a result of more opportunities to analyze input (Bialystok, 1989, 1994). However, statistical analyses showed that the differences in awareness that we found between the short and long exposure groups were due to the fact that learners in the long exposure group had better memory capacities than learners in the short exposure group. Although our results certainly do not go against the views of Cleeremans (2011) and Bialystok (1989, 1994), they cannot provide clear evidence in favor of them. The results underline the already observed links between memory and development of pattern awareness, and also exemplify the importance of assigning participants to experimental groups randomly whenever possible.

Limitations of the present study are similar to the ones already pointed out in the initial study. Although the exposure we used in this study was somewhat longer than the one in the initial study, it can still be considered as relatively short given that it consisted of one session only. We again used retrospective verbal reports to find out if learners became aware of the pattern, and these are known to sometimes lead to overestimating the number of unaware learners (Rebuschat et al., 2015). Another limitation of the study is the use of GJTIs in investigating implicit knowledge. While we are fairly certain that our unaware learners truly were unaware, the task is certainly not ideal for making claims about implicit learning, given that it can trigger learners to search for patterns (Rebuschat et al., 2015). The use of online techniques that do not trigger awareness and do not favor the use of explicit knowledge would be more appropriate for studying implicit learning (Andringa & Curcic, 2015; Andringa & Rebuschat, 2015). Finally, as already indicated, the data samples (i.e., the short and long exposure studies) were collected separately, at different moments, which means that participants were not properly randomly assigned to the long vs. short exposure groups. This later limited us in drawing clear conclusions on the presence or absence of relationship between implicit learning and the development of awareness. In order to further test views of Cleeremans (2011) and Bialystok (1989, 1994), future research would need to randomly assign participants to long vs. short exposure groups, and potentially provide even more exposure to the long exposure group. Also, researchers could use a paradigm that employs more frequent testing and debriefing during exposure in order to see whether learners are more likely to develop awareness after they had accumulated and showed higher levels of implicit knowledge. Although this kind of paradigm may be very challenging to design, it could prove more promising for drawing clearer conclusions on the exact relationship between implicit learning and awareness.
74 Chapter 3
Chapter 4: Effects of input rehearsal on L2 pattern learning*

Abstract

This study investigates whether overt rehearsal of language input may lead to better learning of patterns in adult second language acquisition, which may be expected given that working memory is involved in pattern learning, and rehearsal can help to keep input longer in working memory, according to one of the influential models of working memory (Baddeley & Hitch, 1974). Participants in the study received auditory exposure to a miniature language based on Fijian and a determiner-noun agreement pattern. Learners in the rehearsal group were asked to repeat aloud each phrase or sentence they heard, whereas learners in the silent group heard all phrases and sentences twice. After the exposure, we measured learners’ target pattern knowledge using two grammaticality judgment tasks and a production task. Some learners became aware of the target pattern, whereas the others did not. The results pointed to a paradox: learners who had rehearsed input were less likely to become aware of the pattern than learners who had remained silent; however, the results suggested that once learners became aware of the pattern, rehearsal may have helped them to better apply the pattern they had noticed, and also to better memorize exceptions. For learners who were not aware of the pattern, we found some evidence that rehearsal fostered better memorizing of determiner-noun combinations.

4.1 Introduction

The role of output or language production has often been discussed in the second language acquisition (SLA) literature, and it has been identified as one of the beneficial strategies for language learning, along with exposure to the input (e.g., Izumi, 2002; Izumi & Bigelow, 2000; Swain & Lapkin, 1995). A distinction must be made between superficial production, i.e. mere rehearsal or repetition of the input, and meaningful production, i.e. producing language for communicative purposes. While meaningful production has been widely recognized as essential for language learning (Swain, 1985), the role of

---

*A slightly modified version of this chapter has been submitted as Curcic, M., Andringa, S., & Kuiken, F. (under review). Repetition revisited: Effects of input rehearsal on linguistic pattern learning in second language acquisition.
superficial production has often been neglected (noted in Hulstijn, 2001) or even disputed (e.g., Wong & VanPatten, 2003).

According to one of the influential working memory models (Baddeley & Hitch, 1974), rehearsal is a well-established component of working memory that helps in keeping input longer in short-term memory. Also, numerous studies have found links between working memory capacity on the one hand and vocabulary learning (e.g., Atkins & Baddeley, 1998; Baddeley, Papagno, & Vallar, 1988; Gupta, 2003; Seibert, 1927; Service, 1992; Speciale, Ellis, & Bywater, 2004) and grammar learning (e.g., Denhovska et al., 2015; Kempe, Brooks, & Kharkhurin, 2010; Robinson, 2005b; Sagarra & Herschensohn, 2010) on the other hand. If working memory is involved in language learning, and rehearsal helps to keep input in working memory, it can be expected that rehearsing input may promote language learning via working memory. There is substantial evidence that rehearsal can improve vocabulary learning (Ellis & Beaton, 1993; Papagno et al., 1991; Seibert, 1927; Yoshida & Fukada, 2014). However, the effects of rehearsal on pattern learning have been far less investigated (Hulstijn, 2001), and as a result few solid conclusions can be drawn. The main goal of this study is therefore to shed more light on the effects of rehearsal on learning of grammatical patterns in second language acquisition.

4.1.1 Producing language and its role in SLA
In the late 1970s and early 1980s, language production or output was generally seen as a manifestation of language learning that had already occurred, rather than as an activity that could potentially foster the learning process (e.g., Krashen, 1985). The Comprehensible Output Hypothesis (Swain, 1985) first recognized production as one of the crucial elements of successful language learning, alongside with comprehension. According to this hypothesis, it is meaningful production that fosters acquisition rather than superficial repetition of input. Meaningful production may make learners pay attention to both form and meaning in order to convey a message. Learners get the opportunity to notice the gap between what they want to produce and what they are able to produce, which may lead them to notice the relevant aspects of the language (Swain, 1993, 1995). On the other hand, superficial repetition, rehearsal, or so-called drills require no attention to meaning and also no deep processing of the form. The beneficial role of meaningful production in language learning was later confirmed through extensive experimental research (e.g., Izumi, 2002; Izumi & Bigelow, 2000; Morgan-Short & Bowden, 2006; Swain & Lapkin, 1995). In these studies, learners were required to produce language in meaningful contexts, and such production has been found to foster language learning.
4.1.2 Rehearsal effects as a neglected but promising area of research

One of the neglected but both practically and theoretically relevant questions is how simple rehearsal of language input may influence language learning. This question has received limited attention in the SLA research, possibly due to the fact that rehearsal does not match with the current views of how languages should be thought. As Hulstijn (2001) pointed out, terms such as “drills” and “rehearsal” trigger negative connotations, as they are strongly associated with “superficial parroting of meaningless stimuli”, a practice used in the Audio-Lingual Method of teaching (i.e., teaching method based on behaviorist psychology and used in the 1950s and 1960s). The abandonment of this radical teaching method and the embracement of the communicative approach may likely be the reason why potential beneficial effects of rehearsal in language learning have been neglected, or sometimes even completely dismissed. For instance, Wong and VanPatten (2003) concluded that drills are not only useless, but also potentially harmful. This conclusion was largely based on evidence from comparing the effects of language instruction entirely based on drills with so-called processing instruction (PI). In processing instruction learners are provided with explicit information about the target structure and input processing strategies that can negatively influence their learning of the target structure. Learners are then given structured input activities that encourage them to process the target structure correctly (VanPatten, 2002; Wong & VanPatten, 2003). In such a comparison, it is perhaps not surprising that instruction entirely based on drills would be less effective than any other more communicative or more explicit type of instruction, and very few SLA experts would expect the opposite.

The fact that instruction that is entirely based on drills is not effective does not imply that rehearsal itself is not beneficial, and as Hulstijn (2001) noted, rehearsal practice may have its benefits and should perhaps have a place in language teaching and research. Therefore, more research is needed that uses clear and focused experimental designs to isolate the effects of rehearsal on language learning. Such research may yield important insights in the nature of the language learning process. In addition, it may show what the benefits of rehearsal may be and where its limits lie, and in what way it should or should not be used in language teaching.

4.1.3 The role of rehearsal in short-term memory

From a theoretical point of view, rehearsal could be expected to foster language learning because it enables learners to keep the rehearsed input longer in short-term memory, which may lead to better entrenchment in long term memory. Baddeley and Hitch (1974) proposed a multicomponent model of
working memory, and described it as consisting of three components: the phonological loop, the visuospatial sketchpad, and the central executive. Since then, this model and its later versions (described in Baddeley, 2003, 2012) received wide recognition and confirmation through rich experimental research (e.g., Baddeley & Hitch, 1974; Baddeley, Thomson, & Buchanan, 1975). The model recognizes a role for rehearsal in short-term memory, i.e. it assumes that the phonological loop consists of a storage system and a rehearsal system whose role is to facilitate storage. Experimental evidence for the role of rehearsal in working memory came from word length effect studies. Namely, memorizing five-syllable words has been shown to be more difficult than memorizing monosyllabic words, presumably because longer words are more difficult to rehearse. Importantly, this effect disappears if subvocal rehearsal is prevented by giving learners an extra task of repeating an irrelevant sound (Baddeley et al., 1975).

4.1.4 Effects of rehearsal on vocabulary and pattern learning

There is substantial evidence that short-term memory is involved in L2 vocabulary acquisition (e.g., Atkins & Baddeley, 1998; Baddeley et al., 1988; Gupta, 2003; Seibert, 1927; Service, 1992; Speciale et al., 2004). An early study by Baddeley et al. (1988) showed that a patient with a severe phonological short-term memory deficit failed to learn any novel words in Russian, while having no problems memorizing his native language word pairs. A number of later experimental studies confirmed the role of short-term memory in vocabulary learning, and showed that better short-term memory led to better vocabulary acquisition (e.g., Gupta, 2003; Speciale et al., 2004).

As noted in McDonough and Trofimovich (2016), accumulated experimental evidence has also established a link between working memory capacity and learners’ abilities to become sensitive to linguistic patterns, especially in the domain of agreement morphology. A number of experimental studies showed that better short-term memory leads to better knowledge of L2 patterns (e.g., Denhovska et al., 2015; Kempe et al., 2010; Robinson, 2005b; Sagarra & Herschensohn, 2010).

Given that rehearsal fosters keeping input in short-term memory (Baddeley, 2003, 2012), it can be expected that rehearsing L2 forms – both vocabulary items and grammar structures – may promote their learning (Hulstijn, 2001). Indeed, there are a number of studies that experimentally showed that rehearsal can improve L2 vocabulary acquisition (e.g., Ellis & Beaton, 1993; Papagno et al., 1991; Seibert, 1927; Yoshida & Fukada, 2014). For instance Seibert (1927) showed that French learners retained vocabulary faster and better if they had repeated words aloud, while Papagno et al. (1991) found
that preventing rehearsal by means of articulatory suppression disrupted the ability of Italian adults to learn novel Russian vocabulary.

To our knowledge, only Ellis and Sinclair (1996) investigated the effects of rehearsal on grammar acquisition. They provided some evidence of the beneficial effects of input rehearsal for the acquisition of a soft-mutation morphological pattern in Welsh. Eighty-seven non-Welsh speaking young adults participated in the study. They were auditorily exposed to Welsh words and phrases, and were asked to type in their English translations. The input contained a soft-mutation grammatical pattern, according to which different classes of initial phonemes of nouns undergo different phonological changes. These changes occur in different grammatical contexts and depend on which elements precede the nouns. For instance, initial voiceless plosives p, t, and k change into voiced plosives b, d, and g if a noun is preceded by pronoun his: pen ("head") – ei ben o ("his head"). There were three groups of learners in the study: one group was asked to repeat the input (repetition group), one group remained silent throughout the exposure (silent group), and one group was given an extra counting task to prevent learners from rehearsing subvocally (articulatory suppression group). Prior to the exposure, half of the learners in each group received brief explicit instruction about the target soft-mutation pattern. Learners’ knowledge of the target pattern was assessed using several tests: an oral grammaticality judgment task, a test of metalinguistic knowledge of the rule, and a speech production test. The results showed that rehearsal was beneficial for learning the pattern. Namely, the repetition group outperformed both the suppression group and the silent group. However, Ellis and Sinclair pointed out two important confounds present in the study that made the results inconclusive, and that would need to be addressed in the future research in order to get a clearer picture of the effects of rehearsal. First, they speculated that the suppression group may have performed worse than the repetition group due to the increased cognitive load caused by the additional task. Second, they argued that it remained unclear whether the repetition group performed better than the silent group because learners repeated input aloud or because their vocal repetitions served as extra input. As pointed out by Ellis and Sinclair (1996), it is difficult to say whether the benefits of rehearsal in this study occurred due to output – subjects articulating utterances, or input – additional hearing of their own utterances, or both. The results of Ellis and Sinclair (1996) also showed that learners who rehearsed performed better on the metalinguistic knowledge task compared to both the silent and suppression groups. However, given that half of the learners in each group received explicit rule explanation and half not, it remains unclear if rehearsal promoted rule application of the already aware learners, or it fostered developing metalinguistic knowledge among learners who had not received rule explanation.
Chapter 4

4.1.5 The present study

The main goal of the present study is to investigate the effects of input rehearsal on learning a novel pattern in adult second language acquisition. To our knowledge, only the study by Ellis and Sinclair (1996) has investigated this so far, and we aim to address some of the shortcomings of their study, as pointed out in the study itself and discussed in the previous section. In order to tease apart if the benefits of rehearsal occur due to the production itself or due to the additional input that rehearsal creates, the two groups under comparison in this study (i.e., rehearsal and silent group) received the same amount of input. Also, we introduced no additional cognitively demanding tasks that could disturb learning. In order to see if and how rehearsal affected learners’ awareness of patterns in the input, none of our learners received any explicit instruction, unlike in Ellis and Sinclair (1996).

Participants in the current study were auditorily exposed to a novel miniature language based on Fijian, and the target pattern was determiner-noun agreement. In order to achieve equal exposure between the two groups under comparison, learners in the rehearsal group were asked to repeat every sentence or phrase they heard, whereas learners in the silent group heard every sentence or phrase twice.

Learners’ target structure knowledge was tested using a production task and two oral grammaticality judgment tasks (GJTs): one with items that were familiar from the exposure and one with novel items.

Given that rehearsal is tightly related to working memory, and working memory capacity can influence pattern learning, we also administered two measures of verbal working memory, i.e. the LLAMA D sound recognition task (Meara, 2005), and a non-word repetition task (De Jong & Van der Leij, 1999). Also, given that analytical ability has also been found to affect learning of grammatical patterns (e.g., Erlam, 2005; Harley & Hart, 1997; Robinson, 2005b), we included a measure of learners’ analytical ability – a non-verbal IQ test (Wechsler, 2008). In order to properly isolate the effects of rehearsal on learning, we wanted to make sure that the two groups under comparison did not differ in their working memory capacities and analytical abilities from the start. The design of the study allowed us to also look at whether working memory and analytical ability measures predicted learning of the target pattern, and whether aptitude involvement potentially differed for learners in the silent vs. rehearsal group (i.e., aptitude by treatment interactions). Our goal was to answer the following research questions:

1) Can input rehearsal improve L2 pattern learning? Does it lead to better productive and receptive pattern knowledge, and to heightened awareness of the pattern?
Effects of input rehearsal on L2 pattern learning

2) Can working memory and analytical ability predict L2 pattern learning? Does aptitude involvement differ for different instructional treatments?

We also collected eye-tracking data to investigate individual differences in learners’ online processing of determiners; these results are presented in Chapter 5.

4.2 Method

4.2.1 Participants
Fifty adult native speakers of Dutch participated in the study. They were young adults between 19 and 35 years old ($M_{age} = 23.14, SD = 2.9$), and they were either students or highly educated adults. All participants reported having good hearing, normal or corrected-to-normal vision, no color blindness, and no history of dyslexia. We randomly assigned participants to one of the two experimental groups: the rehearsal group (7 males, 18 females, $M_{age} = 23.76, SD = 3.5$) and the silent group (6 males, 19 females, $M_{age} = 22.52, SD = 2.06$).

The participants reported speaking at least one foreign language, but they were never students of linguistics or graduates in linguistics. Because the target structure may be more easily learnable for participants who speak other foreign languages and languages with similar structures, we wanted to check if learners in the two groups differed in terms of their knowledge of foreign languages. In a questionnaire, participants were asked to list the foreign languages they spoke, and to assess the level of their knowledge (A1 – C2 on the CEFR scales (Council of Europe, 2011)). Based on these reports, we computed two measures that reflected participants’ knowledge of foreign languages: 1) a measure that took into account the number and level of knowledge of all foreign languages (FLK), and 2) a measure that reflected the number and level of foreign languages with a determiner-noun agreement pattern (AgrFLK). A Welch two-sample t-test showed that the rehearsal group ($M = 8.58, SD = 3.29$) and the silent group ($M = 9.86, SD = 3.37$) did not differ significantly on the FLK measure ($t(47.97) = 1.36, p = .18, 95\% CI [-0.61, 3.17]$). A non-parametric Mann-Whitney U test showed that rehearsal ($M = 2.42, SD = 1.99$) and silent learners ($M = 2.52, SD = 2.17$) also did not differ significantly on the AgrFLK measure ($W = 319.5, p = .90, 95\% CI [-1, 1]$).

The experimental session lasted about 1 h and 45 min, and learners received 15 euros for participation. The experiment was approved by the Ethics Committee of the University of Amsterdam.

Whenever t-test assumptions (i.e., normality of distribution and homogeneity of variance) were violated, we used a non-parametric Mann-Whitney U test instead of a t-test.
4.2.2 The target language and the target structure

The target language used in this study was a miniature language that we created by modifying Fijian, which is an Austronesian language of the Malayo-Polynesian family, and is spoken in Fiji. The language consisted of 20 nouns, four verbs, six adjectives, and two determiners (Appendix A). Twenty additional nouns were used for the GJT with novel nouns. All lexical items were based on Fijian, and where possible, both the original form and meaning were retained. However, the original Fijian words frequently had a complex form or they reminded of words from participants’ L1 or potential L2s, in which case we chose a form of another Fijian word.

The language had the same word order (i.e., SVO) as Dutch, and it included a determiner-noun agreement pattern that participants had to acquire. The present tense was marked by a –t suffix in the third person singular, and was created by adding the suffix to the existing Fijian verbs. The nouns were created by adding either –is or –uk to Fijian nouns. All nouns were preceded by one of the two definite determiners: Lep – a masculine determiner or ris – a feminine determiner. The determiner-noun agreement meant that in most cases lep preceded nouns ending in –uk (e.g., lep oseuk (“the ball”)), whereas ris preceded nouns ending in –is (e.g., ris salis (“the chair”)). However, this agreement was not completely regular, since 4 out of 20 nouns that learners were exposed to were exceptions, i.e. the predominant agreement pattern was violated. For instance, the noun touk (“goat”) was preceded by ris instead of lep. We chose to introduce exceptions in the language for two reasons. First, input with exceptions resembles natural input more; and second, we wanted to use exceptions to see if learners would overgeneralize the pattern during the tests, which would be evidence of abstract pattern learning, rather than just memorizing items from the input (Tommasello, 2000). The participants’ L1 – Dutch – also has two definite determiners (de – common and het – neuter), but there is no feminine vs. masculine distinction between them, and also no clear determiner-noun agreement (Booij, 2002).

4.2.3 Language exposure

Learners were not made aware of the purpose of the experiment or of the target structure. They were simply told that they would be learning a new language by looking at images and listening to sentences describing the images. They were also told that their knowledge of the language would be tested from time to time.

Participants did the experiment in a quiet room, and the materials were presented on a computer screen using the E-prime software. The listening materials were presented through loudspeakers. All experimental instructions were written on the computer screen in Dutch, and were also recorded by a
female native speaker of Dutch. We retrieved the images used in the experiment from the Clipart image database (Clipart, Vector Graphics, and Illustrations, 2014), and we then edited them in Photoshop.

*Noun learning and assessment*

The language exposure started with a noun learning phase because we wanted all learners to have perfect and identical knowledge of Fijian nouns before being exposed to the agreement pattern. In this phase, only bare nouns were presented. In every trial, learners saw a simple image and heard the corresponding noun. We asked learners in both groups to repeat every noun they heard aloud, because we wanted the two groups to have similar knowledge of nouns before later receiving the different exposures to the target pattern (i.e., exposure with vs. without rehearsal). Every noun appeared six times during the noun learning phase, and the order of items was identical for all learners.

A noun test followed the noun learning phase. In each trial, learners were presented with four images on the screen, and they heard a noun. We asked them to click on the correct image. Learners received feedback on the accuracy of their response. Each noun was the target noun twice, whereas the distractor images were randomly chosen by the program. The test continued until learners had learned all nouns. The first 40 trials were obligatory and identical for all learners, and they received additional trials with the nouns that they had not recognized correctly. This process continued until learners reached 100% accuracy. We used noun test scores as a measure of learners’ rote memory ability. However, given that the minimum number of trials in the test was fixed (i.e., 40), the noun test scores potentially did not reflect the full range of differences in rote memory between learners. Results of the noun test are presented together with the other aptitude measures in subsection 4.2.7.

*Target structure exposure*

In this part of the exposure, learners heard more complex input that included the determiner-noun agreement target structure (Appendix B). Apart from the 20 nouns that participants had learned during the noun learning phase, this part of the exposure contained some new linguistic elements: three different intransitive verbs, copula *be*, six different adjectives, and two determiners.

In every trial, learners saw an image representing a simple object, intransitive activity, or a state, and they heard a phrase or sentence describing the image. Learners from the rehearsal group heard each phrase or short sentence once, and were asked to repeat it aloud. They were given a few seconds to do this, after which the next trial was presented. During the exposure, all learners from the rehearsal group repeated phrases and sentences
faithfully. The silent group learners heard each phrase or sentence twice, without repeating it. The instruction lasted equally long for the two groups of learners, and both groups received an identical amount of input if we assume that the output of the rehearsal group also served as input. Thus, the two groups differed only in presence or absence of overt rehearsal. In terms of their structure, the trials belonged to one of the following four larger categories:

1) Determiner + noun

Ris salis
"The chair"

2) Determiner + adjective_color + noun

Ris matene salis
"The blue chair"

3) Intransitive sentences: determiner + animate noun + verb_{intransitive}

Ris burogis sisilit.
"The dog is swimming."
4) Intransitive sentences: determiner + inanimate noun + is + adjective

*Ris salis na duka.*
“*The chair is broken.*”

Learners received a total of 308 trials, divided into two parts with a short break in the middle. The first part consisted of 176 trials and lasted about 18 minutes, whereas the second part consisted of 132 trials, and lasted about 12 minutes. During the exposure, 18 out of 20 nouns appeared 14 times, while two regular nouns (feminine noun *burogis* – “dog” and masculine noun *dawauk* – “cat”) were overrepresented and appeared 28 times each. We chose to present two nouns more frequently so that the input would resemble the natural input more (Ellis & Ferreira-Junior, 2009). The trials were presented in the same order to all learners.

Learners received equal exposure to both *lep* – *uk* and *ris* – *is* agreement patterns. The two determiner categories had the same number of overrepresented vs. underrepresented nouns, animate vs. inanimate nouns, non-biological vs. biological gender nouns (i.e., boy, girl, man, woman).

Four out of 20 nouns – two feminine and two masculine – were exceptions to the dominant determiner-noun agreement pattern, and exception items were gradually introduced after 30 trials.

### 4.2.4 Production test

After the exposure, learners in both groups did a production test in order to measure their productive knowledge of determiner-noun combinations and the agreement pattern. The test consisted of 40 trials. In each trial, learners saw an image on the screen representing a state or an intransitive activity, and they were asked to produce orally the corresponding sentence in the new language. In each trial, learners were expected to produce the correct determiner-noun-verb combination based on the image they saw. Each of the 20 nouns appeared twice as the target, and all images in the production test trials were familiar to learners from the exposure phase.

### 4.2.5 Oral GJTs

After the production test, we administered two GJTs to test learners’ receptive knowledge of the target structure. The first GJT contained novel nouns that had
not appeared in the exposure phase (Appendix C), whereas the second one featured familiar nouns that learners had been exposed to (Appendix D).

Each trial consisted of a simple image presented on the screen, and a short phrase consisting of a determiner and a noun that participants heard. At the beginning of the GJT with novel nouns, learners were told that they would hear new phrases that they had not heard before, whereas at the beginning of the GJT with familiar nouns, they were told that the phrases would be familiar to them. For each phrase they heard, they were asked to decide if it was good or not in the language they had learned, and they could re-play the phrase as many times as they wanted. We encouraged learners to use their intuition or to guess if they were uncertain about the correct answer. For all learners, the test items were presented in the same order.

In both GJTs, we systematically varied noun type – *is* nouns vs. *-uk* nouns, and the grammaticality of the phrase – grammatical or ungrammatical. This means that there were four conditions, with 10 items per condition. Ungrammaticality of the phrases always stemmed from incorrect determiner-noun combinations.

The GJT with familiar nouns featured exceptions as well, and accuracy was always scored according to what learners had heard in the exposure phase.

### 4.2.6 Debriefing

After the language exposure and tests, we debriefed learners using a protocol to check if they had consciously noticed the target pattern, and if so, in which stage of the experiment (i.e., during the exposure, the production test, GJTs, or during debriefing). We also asked learners about what kind of knowledge they relied on during each of the two GJTs. We considered learners to have been aware of the pattern if they could verbalize it at least partially. If learners did not report any pattern awareness, we told them that there was a pattern in the language and gave them the opportunity to guess what it was. If still no awareness was shown, we told them what the pattern was, and that the language was a non-existing language based on Fijian. The debriefing session helped us classify learners as aware or unaware of the target pattern in different stages of the experiment.

### 4.2.7 Aptitude measures

We administered several aptitude tests in order to measure learners’ memory and analytical abilities: the LLAMA D sound recognition task (Meara, 2005), a non-word repetition task (NWR; De Jong & Van der Leij, 1999) – measures of verbal working memory, the IQ test (Wechsler, 2008) – a non-verbal measure of analytical ability, and the noun test – a measure of rote memory. Below we present the tests in the order in which they were administered.
The noun test
On average, learners needed 45.26 items ($SD = 6.63$, $min = 40$, $max = 76$) to pass the test. Note that on this test, lower performance scores indicate that participants were quicker to learn. There were 9 learners who scored 40, which means that for those learners, the test potentially did not measure the full range of rote memory abilities. According to a Mann-Whitney U test, learners in the rehearsal ($M = 46.32$, $SD = 5.94$) and silent group ($M = 44.2$, $SD = 7.22$) did not differ significantly in their performance on the noun test ($W = 218.5$, $p = .07$, $95\%$ CI $[-4, 0]$).

The non-verbal IQ test
We administered the Matrix Reasoning subtest of the Wechsler Adult Intelligence Scale – 4th edition (WAIS-IV) (Wechsler, 2008) as a measure of learners’ analytical ability. The test measures non-verbal abstract problem solving and inductive reasoning ability. There were 26 items in the test, and the learners’ task was to choose the missing part of a visual pattern. There were always five options to choose from. The test was administered after the first part of the exposure to the target structure, and participants were given maximum 20 minutes to complete the test.

The average accuracy on the test was 23.32 ($SD = 2.08$, $min = 14$, $max = 26$). We ran a non-parametric Mann-Whitney U test to check if learners in the rehearsal group ($M = 22.88$, $SD = 2.57$) differed in their performance on the IQ test from learners in the silent group ($M = 23.76$, $SD = 1.36$). According to the test, the two groups did not differ significantly in their IQ scores ($W = 367.5$, $p = .27$, $95\%$ CI $[0, 1]$).

The LLAMA D sound recognition task
The LLAMA D task is a part of the LLAMA language aptitude test battery (Meara, 2005), and was used to assess learners’ verbal working memory. The test measures the ability to recognize spoken language that you have been exposed to before. According to Meara (2005), this aptitude is involved in the learning of vocabulary and morphological patterns.

Participants first listened to a sequence of 10 words in an unknown language, after which they did a 30-item test, in which they heard both previously heard words and novel words. Learners’ task was to decide for each word whether they had heard it before. If they decided correctly, they gained points, whereas if they decided wrongly, they lost points. The maximum score was 75, and the task lasted about 5 minutes. It was administered after the debriefing session.
Participants had an average accuracy of 28.1 (SD = 12.73, min = 0, max = 55). We ran a Welch two-sample t-test, and found no significant differences between the rehearsal (M = 30, SD = 13.46) and silent group (M = 26.2, SD = 11.93) in their performance on LLAMA D (t(47.31) = -1.06, p = .30, 95% CI [-11.04, 3.44]).

The non-word repetition task (NWR)
The non-word repetition task (De Jong & Van der Leij, 1999) was used to assess learners’ verbal working memory. The test consisted of two practice trials and 40 experimental trials recorded by a female native speaker of Dutch. In each trial, participants heard a non-word, and were asked to repeat it aloud as accurately as possible. Non-words consisted of 2 to 5 syllables, and featured between 5 and 15 phonemes.

Scoring was done by the first author of this study; all omissions, substitutions, or insertions of additional phonemes were counted as errors. For each item, we calculated the percentage of phonemes correctly repeated since this method is generally more sensitive than calculating the number of correctly repeated items (Graf Estes, Evans, & Else-Quest, 2007; Rispens, Baker, & Duinmeijer, 2015). Twenty percent of the data was randomly chosen, and was analyzed by another trained researcher. The inter-rater reliability was excellent, as indicated by the intraclass correlation coefficient of 0.96 (95% CI [0.94, 0.97]). Due to a technical error, half of the data of one participant from the silent group were not recorded, so this participant’s scores were not taken into account in further analyses that involved the NWR task.

On average, participants correctly repeated 93.05% of phonemes (SD = 3.39, min = 82.54, max = 98.37). A non-parametric Mann-Whitney U test showed that there was no significant difference between the rehearsal (M = 92.72, SD = 3.98) and the silent group (M = 93.4, SD = 2.67) in their performance on the NWR task (W = 308, p = .88, 95% CI [-1.71, 2.12]). The NWR task was administered after LLAMA D.

4.2.8 Statistical procedures
We analyzed the experimental results using R (R Development Core Team, 2011). Prior to analyses, we centered all continuous variables (e.g., aptitude measure scores). This means that each learner’s scores showed the deviation from the mean score of all participants. For all categorical variables, we specified explicit contrasts, where we assigned -0.5 to one level of the variable and +0.5 to the other level of the variable.

To analyze learners’ performance on the GJT, the production test, and the relation between aptitude measures and GJT or production, we used generalized linear mixed models (GLMMs), which include both fixed and
random effects. The dependent variable was learners’ accuracy, a binomial variable coded as 1 (correct) or 0 (incorrect). We always used maximally specified models, which means that we included all fixed effects under investigation, and all two-way and three-way interactions between them.

Since Barr et al. (2013) recommend applying maximal random effects structures when using linear mixed-effects models for confirmatory hypothesis testing, we always included both by-subject and by-item random intercepts and slopes for each fixed effect included in the model. Whenever subjects or items were nested within a particular fixed effect variable, we did not include the corresponding by-subject or by-item random slopes because such models would be unidentifiable. For instance, given that subjects were nested within group (i.e., rehearsal vs. silent group), we included the by-item but not the by-subject random slope for the group fixed effect.

For the analysis of relation between aptitude measures and learners’ awareness of the target pattern, we used generalized linear models (GLMs), which include only random effects.

4.3 Results

Learners in this study differed substantially in terms of whether they became aware of the target pattern, and we needed to take these differences into consideration in our data analyses. In this section, we first present awareness-related results. We will then consider the results of the GJT with familiar nouns in order to look at learners’ performance based on memory, and potentially also pattern knowledge. The GJT with novel nouns will allow us to look only at learners’ pattern knowledge. Finally, the production task results will give us insights in learners’ productive knowledge of determiner-noun combinations, based on memory, and also potentially pattern knowledge.

4.3.1 Debriefing results – awareness

During the debriefing, most learners reported that they focused on learning nouns, verbs, and adjectives during the exposure, and paid much less attention to the other aspects of the language. Most learners – 41 out of 50 – became aware at some point that lep was a masculine determiner, whereas ris was a feminine determiner. Twenty-three learners were not only aware of the gender distinction, but they also noticed the agreement between the determiner and the noun during or immediately after the experiment, and they were able to reproduce it correctly. Table 4.1 summarizes the results. As the experiment progressed, more learners in both groups developed awareness of the pattern.

During debriefing, learners reported having had different sources of reliance during the GJTs: memory, guessing, intuition, the target pattern, a wrong pattern, and some learners reported having judged items based on
nouns only, which we labeled as *wrong focus*. We did not take the differences in sources of reliance into account in our analyses. Also, we decided not to exclude learners who had a wrong focus or relied on a wrong pattern because these sources of reliance may have resulted from learners’ insensitivity to the target pattern.

In order to find out if learners in the two groups differed in their awareness of the target pattern, we built a generalized linear model (GLM) with awareness as a dependent variable and group as a fixed effect. There was a strong main effect of group, such that learners who had rehearsed input were less likely to become aware of the target pattern than learners who had not ($b = -1.52, SE = 0.61, p = .01$, 95% CI [-2.74, -0.3], $OR = 0.22$). In the subsequent analyses of the two GJTs and the production test scores, we always considered aware and unaware learners separately.

### Table 4.1. Overview of the cumulative number of learners aware of the target pattern at different stages in the experiment, split by group

<table>
<thead>
<tr>
<th>Stage in the experiment</th>
<th>Rehearsal group</th>
<th>Silent group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td>5/25</td>
<td>9/25</td>
</tr>
<tr>
<td>Production test</td>
<td>5/25</td>
<td>9/25</td>
</tr>
<tr>
<td>GJT novel nouns</td>
<td>6/25</td>
<td>10/25</td>
</tr>
<tr>
<td>GJT familiar nouns</td>
<td>7/25</td>
<td>11/25</td>
</tr>
<tr>
<td>After experiment</td>
<td>7/25</td>
<td>16/25</td>
</tr>
</tbody>
</table>

#### 4.3.2 GJT with familiar nouns

We analyzed the results of the GJT with familiar nouns in order to find out if rehearsal promoted learners’ knowledge of the items that they had heard in the exposure. The task consisted both of regular items ($n = 32$) and exception items ($n = 8$). Accuracy was based on what learners had heard in the exposure. Eight of the 32 regular items had biological gender, and we excluded them from the analysis because knowledge of gender only – i.e. that *ris* is a feminine determiner, whereas *lep* is a masculine determiner – could have led to the correct response on these items. A Mann Whitney U test showed that indeed learners had higher scores on biological regular items ($M = 87.75, SD = 19.96$) than on non-biological regular items ($M = 68.83, SD = 20.35, W = 1957, p < .001$, 95% CI [0.08, 0.33]). We analyzed the remaining regular and exception items together, but we included regularity as a variable. We hypothesized that if
learners had acquired the underlying pattern, and did not just memorize the
determiner-noun combinations from the input, they would overgeneralize the
pattern to exceptions, and as a result, they would be more accurate on regular
items than on exceptions.

The average accuracy on the test was 64.75% ($SD = 16.11$). Aware and
unaware learners were found to perform very differently. Learners who were
aware of the pattern ($n = 18$) scored 76.39% correct ($SD = 12.87$), whereas the
learners who reported no awareness of the pattern ($n = 32$) scored 58.2%
correct ($SD = 14.01$). Because of this, data of aware and unaware learners were
analyzed separately in subsequent analyses.

For the analysis of aware learners’ performance, we built a model with the
following fixed effects: group (rehearsal vs. silent), grammaticality (grammatical vs. non-grammatical items), and regularity (regular items vs.
exceptions). The main effect of group was not statistically significant although
there was a tendency that learners in the rehearsal group performed better
than learners in the silent group ($b = 1.13, SE = 0.59, p = .056, 95\% CI [-0.05,
2.31], OR = 3.09$). There was a main effect of grammaticality – learners were
less accurate on ungrammatical items ($b = -1.11, SE = 0.37, p = .003, 95\% CI [-
1.85, -0.38], OR = 0.33$), meaning that they had a bias towards marking items as
grammatical. There was also a main effect of regularity, meaning that learners
were less accurate on exceptions than on regular items ($b = -3.02, SE = 0.81, p <
.001, 95\% CI [-4.63, -1.41], OR = 0.05$). We also found a significant interaction
between group and grammaticality (Table 4.2), such that learners from the
rehearsal group performed much better on ungrammatical items compared to
learners from the silent group ($b = 1.79, SE = 0.68, p = .009, 95\% CI [0.42, 3.15],
OR = 5.97$). The other two-way interactions were not statistically significant:

- group x regularity ($b = 0.46, SE = 1.59, p = .77, 95\% CI [-2.71, 3.64], OR = 1.59$)
- regularity x grammaticality ($b = 0.47, SE = 0.7, p = .51, 95\% CI [-0.94, 1.87],
$ OR = 1.59$).

The three-way interaction between group, regularity, and grammaticality was not statistically significant either ($b = 1.36, SE = 1.29, p =
.29, 95\% CI [-1.22, 3.94], OR = 3.89$).

These results confirm that aware learners indeed acquired the abstract
underlying pattern: they were less accurate on exceptions than on regular
items, and this means that they overgeneralized the pattern to exceptions,
treating exceptions as regular items. Also, rehearsal seems to have helped
aware learners to better apply the pattern they had noticed, and also to
memorize exceptions better, i.e. aware learners who had rehearsed input were
able to identify ungrammatical items – both regular items and exceptions –
with much more ease compared to the learners who had remained silent during
the exposure (Table 4.2).
Table 4.2. Overview of aware learners’ mean accuracies and their standard deviations (SD) on the GJT with familiar nouns, split by group, regularity, and grammaticality of items

<table>
<thead>
<tr>
<th></th>
<th>Regular items</th>
<th>Exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rehearsal group</td>
<td>95.24 (6.56)</td>
<td>91.67 (9.62)</td>
</tr>
<tr>
<td>(n = 7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silent group</td>
<td>91.67 (14.91)</td>
<td>73.49 (22.61)</td>
</tr>
<tr>
<td>(n = 11)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Next, we analyzed the data of learners who reported having no awareness of the target pattern during the test (n = 32). For these learners, we found a tendency towards overall above chance performance (b = 0.49, SE = 0.27, p = .07, 95% CI [-0.05, 1.05], OR = 1.63). We then built a model that included group, reliability, and grammaticality as fixed effects. There were no main effects of group (b = 0.2, SE = 0.3, p = .5, 95% CI [-0.4, 0.8], OR = 1.22), and reliability (b = -0.13, SE = 0.33, p = .69, 95% CI [-0.78, 0.52], OR = 0.88), but there was a main effect of grammaticality, implying that learners were less accurate on ungrammatical items (b = -2.38, SE = 0.48, p < .001, 95% CI [-3.33, -1.43], OR = 0.09). We found no significant two-way interaction between group and grammaticality (b = 1.14, SE = 0.81, p = .16, 95% CI [-0.47, 2.75], OR = 3.13), although there was a tendency that learners in the rehearsal group were more accurate on ungrammatical items than learners from the silent group. The other two-way interactions were not statistically significant either: group x regularity (b = -0.02, SE = 0.41, p = .96, 95% CI [-0.84, 0.8], OR = 0.98) and regularity x grammaticality (b = 0.55, SE = 0.62, p = .38, 95% CI [-0.7, 1.8], OR = 1.74).

Importantly, we found a three-way interaction between group, grammaticality, and reliability (b = 1.97, SE = 0.74, p = .008, 95% CI [0.5, 3.45], OR = 7.18), indicating that the group x grammaticality interaction was present in the exceptions more than in the regular items. Judging from the descriptives (Table 4.3), this group x grammaticality interaction in the exceptions may be driven mainly by poorer performance on the grammatical exceptions of the rehearsal group, and better performance of the rehearsal group on the ungrammatical exceptions.
Table 4.3. Overview of unaware learners’ mean accuracies and their standard deviations (SD) on the GJT with familiar nouns, split by group, regularity, and grammaticality of items

<table>
<thead>
<tr>
<th></th>
<th>Regular items</th>
<th>Exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gramm. (SD)</td>
<td>Ungramm. (SD)</td>
</tr>
<tr>
<td>Rehearsal group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 18)</td>
<td>81.94 (14.08)</td>
<td>38.89 (30.38)</td>
</tr>
<tr>
<td>Silent group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 14)</td>
<td>80.36 (17.48)</td>
<td>32.74 (26.65)</td>
</tr>
</tbody>
</table>

If learners acquired the abstract pattern, we expected them to perform better on regular items than on exceptions. This seems to have been the case for the rehearsal group, but only on the grammatical items, whereas on the ungrammatical items, there was a trend in the opposite direction (Table 4.3). This does not give us sufficient evidence to claim that the rehearsal group showed pattern knowledge without awareness. However, the results do indicate that rehearsal may have had some beneficial effects on learners’ acquisition of determiner-noun combinations in that learners who had rehearsed input scored better on ungrammatical items than learners who had not rehearsed, and this was especially visible on ungrammatical exceptions.

4.3.3 GJT with novel nouns
The goal of the GJT with novel nouns was to see if learners could generalize the target pattern to novel nouns, and if their ability to do so was influenced by the presence or absence of rehearsal during the exposure.

The average accuracy on the test was 63.15% (SD = 19.67). Learners who were aware of the pattern scored 88.13% (SD = 13.43), whereas the learners who reported no awareness of the pattern scored 51.4% (SD = 6.91).

We first analyzed the data of aware learners (n = 16). We built a model with group and grammaticality as fixed effects. We found no main effects of group (b = 0.25, SE = 0.78, p = .75, 95% CI [-1.31, 1.81], OR = 1.28) and grammaticality (b = -0.57, SE = 0.75, p = .45, 95% CI [-2.07, 0.94], OR = 0.57). The interaction between group and grammaticality was also not significant (Table 4.4) (b = 0.85, SE = 1.29, p = .51, 95% CI [-1.73, 3.44], OR = 3.35). As these results suggest, we found no evidence that rehearsal fostered better performance on the GJT with novel nouns for learners aware of the pattern.
Table 4.4. Overview of aware learners’ mean accuracies and their standard deviations (SD) on the GJT with novel nouns, split by group and grammaticality of items

<table>
<thead>
<tr>
<th>Grammatical items</th>
<th>Ungrammatical items</th>
</tr>
</thead>
<tbody>
<tr>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
</tr>
<tr>
<td>Rehearsal group</td>
<td>Silent group</td>
</tr>
<tr>
<td>(n = 6)</td>
<td>(n = 10)</td>
</tr>
<tr>
<td>93.33 (6.83)</td>
<td>95 (4.71)</td>
</tr>
<tr>
<td>87.5 (8.8)</td>
<td>78.5 (31.45)</td>
</tr>
</tbody>
</table>

We found that learners without pattern awareness (n = 34) did not perform significantly above chance on the GJT with novel items (b = 0.06, SE = 0.11, p = .57, 95% CI [-0.15, 0.27]). We built a model with group (rehearsal vs. silent) and grammaticality of items (grammatical vs. ungrammatical) as fixed effects. We found no main effect of group (b = 0.02, SE = 0.13, p = .9, 95% CI [-0.24, 0.27], OR = 1.02), but there was a main effect of grammaticality (b = -1.3, SE = 0.33, p < .001, 95% CI [-1.96, -0.65], OR = 0.27), indicating that learners were biased towards marking items as grammatical (Table 4.5). The interaction between group and grammaticality was not significant (b = 0.09, SE = 0.66, p = .89, 95% CI [-1.23, 1.41], OR = 1.09).

Table 4.5. Overview of unaware learners’ mean accuracies and their standard deviations (SD) on the GJT with novel nouns, split by group and grammaticality of items

<table>
<thead>
<tr>
<th>Grammatical items</th>
<th>Ungrammatical items</th>
</tr>
</thead>
<tbody>
<tr>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
</tr>
<tr>
<td>Rehearsal group</td>
<td>Silent group</td>
</tr>
<tr>
<td>(n = 19)</td>
<td>(n = 15)</td>
</tr>
<tr>
<td>64.47 (16.32)</td>
<td>64.67 (19.04)</td>
</tr>
<tr>
<td>38.68 (19.06)</td>
<td>37.67 (23.14)</td>
</tr>
</tbody>
</table>

Finally, we wanted to check if there were any learners unaware of the pattern who performed significantly above chance on the test. According to the binomial test, learners who scored 67.5% or more could be reliably classified as having performed above chance (p < 0.05). There was only one such learner. He was in the rehearsal group, and had an accuracy of 70%. He reported having used a wrong rule as a basis for doing the test.
To sum up, in the GJT with novel nouns, learning was found only among aware learners, whereas learners unaware of the pattern did not perform above chance. We found no clear evidence that rehearsal promoted performance on the GJT with novel nouns for either learners who reported awareness of the pattern, or those who did not. However, if we look at the descriptive statistics in Table 4.4 above, we can observe that learners who had rehearsed had higher accuracies on ungrammatical items than learners who had not rehearsed. It is possible that we could not statistically detect this difference due to the small number of learners in the groups under comparison. Also, judging from the standard deviations (Table 4.4), silent learners showed substantial individual differences in their performance on ungrammatical items, whereas these differences were much smaller in the rehearsal group. This may imply that rehearsal had some effect on diminishing individual differences between learners, i.e. learners who had rehearsed were more consistently accurate than learners who had remained silent. However, these observations need to be treated with caution, and are in need of replication with bigger samples.

4.3.4 Production test
Learners’ responses on the production test were recorded, transcribed, and coded according to several established criteria and categories. The coding was done by the first author of the study. Each item was scored as 1 or 0 on the following categories: producing the target or the opposite determiner recognizably (lep vs. ris), producing the target noun recognizably, and producing the target or the opposite noun ending (-uk vs. -is). Scores on these categories made it possible to evaluate learners’ production of determiner-noun combinations and also determiner-noun agreement. Twenty percent of the data was randomly chosen and analyzed by another trained researcher. The interrater reliability on all the relevant categories later used in the analyses was excellent, and ranged between 0.86 and 0.94.

We first analyzed learners’ production of regular items. We were interested in whether rehearsal promoted learners’ production of determiner-noun agreement, but the analysis of regular items could not tell us if this was based on memorizing or pattern knowledge. Learners’ production on each item was scored as correct if learners produced the target noun together with the target determiner and noun ending. Learners who were aware of the pattern during the test \((n = 14)\) scored 57.59% \((SD = 23.41)\), whereas learners unaware of the pattern \((n = 36)\) scored 28.3% \((SD = 23.49)\). We built separate models for aware and unaware learners, and included group as a fixed effect in each model. We found no evidence that rehearsal improved production of regular items (Table 4.6) in either aware learners \((b = 0.98, SE = 0.81, p = .23, 95\% CI [-0.64, 2.59])\),
OR = 2.65), or unaware learners (b = -0.76, SE = 0.76, p = .31, 95% CI [-2.28, 0.75], OR = 0.47). However, again, it is worth noting that among aware learners, the rehearsal group scored higher and had lower standard deviation than the silent group (Table 4.6). It is possible that this difference did not reach significance due to lack of power. There were only 5 aware learners in the rehearsal group and 9 aware learners in the silent group.

Table 4.6. Overview of learners’ mean accuracies and their standard deviations (SD) on the production test, split by group and pattern awareness

<table>
<thead>
<tr>
<th></th>
<th>Aware learners</th>
<th>Unaware learners</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% correct (SD)</td>
<td>% correct (SD)</td>
</tr>
<tr>
<td><strong>Rehearsal group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>67.5 (16.33)</td>
<td>25.94 (23.95)</td>
</tr>
<tr>
<td></td>
<td>(n = 5)</td>
<td>(n = 20)</td>
</tr>
<tr>
<td><strong>Silent group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>52.08 (25.72)</td>
<td>31.25 (23.33)</td>
</tr>
<tr>
<td></td>
<td>(n = 9)</td>
<td>(n = 16)</td>
</tr>
</tbody>
</table>

Next, we analyzed learners’ production of exceptions, which was scored in two different ways: whether learners’ production of exceptions agreed with how exceptions were presented in the input (memory-based) or whether learners’ production of exceptions agreed with the dominant pattern (pattern-based). By scoring whether learners produced exceptions as such, or produced them according to the pattern, it was possible to disentangle if learners produced based on memory (if they could correctly reproduce exceptions) or based on the dominant pattern (if they produced overgeneralized versions of exceptions). For learners aware of the pattern, this analysis allowed us to see if rehearsal fostered better memorizing of exceptions and led to less overgeneralization. For unaware learners on the other hand, this analysis could give us insight in whether learners relied on memory or the underlying pattern in their production, and whether this differed for the rehearsal and silent groups.

We first analyzed the results of aware learners. We ran separate models for memory-based and pattern-based production, and we always included group as the fixed effect. Aware learners who had rehearsed were better at memory-based production than aware learners who had not rehearsed (Table 4.7 – memory-based production), i.e. they could better memorize and produce exceptions the way they had heard them (b = 2.23, SE = 0.95, p = .02, 95% CI [0.32, 4.14], OR = 9.3). Similarly, aware learners who had rehearsed tended to produce fewer overgeneralized exceptions than silent learners (Table 4.7 – pattern-based production), but this difference did not reach statistical
significance, possibly due to the lack of power ($b = -2.92, SE = 2.12, p = .17, 95\% CI [-7.16, 1.33], OR = 0.05$).

**Table 4.7. Overview of aware learners' mean accuracies and their standard deviations (SD) on the production test, split by group and memory-based vs. pattern-based performance**

<table>
<thead>
<tr>
<th></th>
<th>Memory-based production</th>
<th>Pattern-based production</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% produced (SD)</td>
<td>% produced (SD)</td>
</tr>
<tr>
<td>Rehearsal group</td>
<td>65 (29.84)</td>
<td>5 (11.18)</td>
</tr>
<tr>
<td>(n = 5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silent group</td>
<td>29.17 (25.77)</td>
<td>22.22 (20.52)</td>
</tr>
<tr>
<td>(n = 9)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Turning to the production of unaware learners, we again ran separate models for memory-based and pattern-based production, and we included group as the fixed effect. There was no evidence that silent and rehearsal learners differed in their memory-based production of exceptions ($b = 0.04, SE = 0.6, p = .94, 95\% CI [-1.17, 1.25], OR = 1.04$), but learners who had rehearsed produced overgeneralized exceptions less often than silent learners (Table 4.8 – pattern-based production), and this effect was statistically significant ($b = -0.94, SE = 0.48, p = .049, 95\% CI [-1.89, 0.02], OR = 0.39$).

**Table 4.8. Overview of unaware learners' mean accuracies and their standard deviations (SD) on the production test, split by group and memory-based vs. pattern-based performance**

<table>
<thead>
<tr>
<th></th>
<th>Memory-based production</th>
<th>Pattern-based production</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% produced (SD)</td>
<td>% produced (SD)</td>
</tr>
<tr>
<td>Rehearsal group</td>
<td>19.38 (22.39)</td>
<td>13.13 (13.74)</td>
</tr>
<tr>
<td>(n = 20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silent group</td>
<td>21.88 (17.38)</td>
<td>24.22 (20.14)</td>
</tr>
<tr>
<td>(n = 16)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.3.5 The role of working memory, rote memory, and analytical ability

We were interested in whether working memory (measured by LLAMA D and NWR task), rote memory (measured by the noun test), and analytical ability (measured by the IQ test) predicted learners’ performance on the two GJT's and the production task. The design of the study also allowed us to check if aptitude involvement differed for the two instruction groups. Finally, we looked at whether any of the aptitude measures could predict if learners would become aware of the pattern. Since we were interested in whether each of the aptitude measures could predict learning, we ran separate models for all aptitude measures we investigated.

First, we checked correlations between different aptitude measures, and only the analytical ability measure (i.e., IQ) and the noun test scores correlated significantly \((r = -0.31, p = .03)\), implying that when learners scored higher on the analytical ability test, they needed fewer trials to pass the noun test.

We found that LLAMA D, noun test scores, and IQ scores significantly predicted learners’ performance on the GJT with familiar nouns (Table 4.9; Figure 4.1). When learners performed better on these measures, they were also more accurate on the GJT with familiar nouns. There were no treatment (silent vs. rehearsal) by aptitude interactions (Table 4.9).
### Table 4.9. Overview of aptitude involvement in learners’ performance on the GJT with familiar nouns, GJT with novel nouns, and learners’ pattern awareness

<table>
<thead>
<tr>
<th>GJT WITH FAMILIAR NOUNS</th>
<th>b</th>
<th>SE</th>
<th>p</th>
<th>95% CI</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main effects of aptitude measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D</td>
<td>0.03</td>
<td>0.01</td>
<td>.02</td>
<td>[0.004, 0.06]</td>
<td>1.03</td>
</tr>
<tr>
<td>Noun test</td>
<td>-0.05</td>
<td>0.02</td>
<td>.001</td>
<td>[-0.09, -0.02]</td>
<td>0.95</td>
</tr>
<tr>
<td>IQ</td>
<td>0.11</td>
<td>0.04</td>
<td>.008</td>
<td>[0.03, 0.2]</td>
<td>1.12</td>
</tr>
<tr>
<td>NWR</td>
<td>0.07</td>
<td>0.05</td>
<td>.19</td>
<td>[-0.04, 0.17]</td>
<td>1.07</td>
</tr>
<tr>
<td><strong>Aptitude x group interactions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D x group</td>
<td>0.01</td>
<td>0.03</td>
<td>.79</td>
<td>[-0.05, 0.06]</td>
<td>1.01</td>
</tr>
<tr>
<td>Noun test x group</td>
<td>-0.05</td>
<td>0.03</td>
<td>.09</td>
<td>[-0.11, 0.01]</td>
<td>0.95</td>
</tr>
<tr>
<td>IQ x group</td>
<td>0.2</td>
<td>0.14</td>
<td>.14</td>
<td>[-0.07, 0.48]</td>
<td>1.23</td>
</tr>
<tr>
<td>NWR x group</td>
<td>-0.12</td>
<td>0.08</td>
<td>.12</td>
<td>[-0.28, 0.04]</td>
<td>0.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GJT WITH NOVEL NOUNS</th>
<th>b</th>
<th>SE</th>
<th>p</th>
<th>95% CI</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main effects of aptitude measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D</td>
<td>-0.001</td>
<td>0.02</td>
<td>.96</td>
<td>[-0.03, 0.03]</td>
<td>0.999</td>
</tr>
<tr>
<td>Noun test</td>
<td>-0.06</td>
<td>0.02</td>
<td>.02</td>
<td>[-0.1, -0.01]</td>
<td>0.95</td>
</tr>
<tr>
<td>IQ</td>
<td>0.07</td>
<td>0.04</td>
<td>.1</td>
<td>[-0.02, 0.16]</td>
<td>1.07</td>
</tr>
<tr>
<td>NWR</td>
<td>0.1</td>
<td>0.03</td>
<td>&lt; .001</td>
<td>[0.04, 0.16]</td>
<td>1.11</td>
</tr>
<tr>
<td><strong>Aptitude x group interactions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D x group</td>
<td>-0.005</td>
<td>0.03</td>
<td>.87</td>
<td>[-0.06, 0.05]</td>
<td>0.995</td>
</tr>
<tr>
<td>Noun test x group</td>
<td>-0.01</td>
<td>0.04</td>
<td>.77</td>
<td>[-0.1, 0.07]</td>
<td>0.99</td>
</tr>
<tr>
<td>IQ x group</td>
<td>-0.02</td>
<td>0.17</td>
<td>.91</td>
<td>[-0.36, 0.32]</td>
<td>0.98</td>
</tr>
<tr>
<td>NWR x group</td>
<td>0.02</td>
<td>0.1</td>
<td>.85</td>
<td>[-0.17, 0.21]</td>
<td>1.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PATTERN AWARENESS</th>
<th>b</th>
<th>SE</th>
<th>p</th>
<th>95% CI</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main effects of aptitude measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D</td>
<td>0.02</td>
<td>0.02</td>
<td>.33</td>
<td>[-0.02, 0.07]</td>
<td>1.02</td>
</tr>
<tr>
<td>Noun test</td>
<td>-0.2</td>
<td>0.1</td>
<td>.02</td>
<td>[-0.39, -0.05]</td>
<td>0.82</td>
</tr>
<tr>
<td>IQ</td>
<td>0.14</td>
<td>0.15</td>
<td>.37</td>
<td>[-0.14, 0.48]</td>
<td>1.15</td>
</tr>
<tr>
<td>NWR</td>
<td>0.17</td>
<td>0.1</td>
<td>.08</td>
<td>[-0.01, 0.39]</td>
<td>1.19</td>
</tr>
</tbody>
</table>
When it comes to the GJT with novel nouns, the noun test scores and the NWR test scores significantly predicted learners’ accuracy (Table 4.9; Figure 4.2). As in the GJT with familiar nouns, no aptitude by treatment interactions were found.
We were also interested in whether learners’ awareness of the target pattern could be predicted by any of the aptitude measures, and we found that only learners’ performance on the noun test predicted their awareness of the target pattern, with learners who performed better on the noun test being more likely to also notice the target pattern (Table 4.9).

As for the production task, IQ and noun test scores significantly predicted learners’ production of regular items (Table 4.10), implying that learners who had better performance on the IQ and noun test were also more likely to correctly produce regular items. These effects did not differ for the two groups of learners (Table 4.10).

The NWR task performance predicted learners’ memory-based production of exceptions (Table 4.10). There was also an interaction between IQ scores and group, such that higher IQ scores were related to more memory-based production of exceptions in the rehearsal but not in the silent group (Table 4.10).

None of the aptitude measures significantly predicted learners’ production of overgeneralized exceptions. There was an IQ x group interaction (Table 4.10) – higher IQ scores led to more production of overgeneralized exceptions in the silent learners compared to the learners who had rehearsed. The reason why IQ scores were less predictive in the rehearsal group may be that these learners managed to memorize exceptions thanks to rehearsal, so they produced overgeneralized exceptions less frequently.
Table 4.10. Overview of aptitude involvement in learners’ performance on the production task

<table>
<thead>
<tr>
<th>PRODUCTION TASK – regular items</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Main effects of aptitude measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLAMA D</td>
<td>0.02</td>
<td>0.03</td>
<td>0.47</td>
<td>[-0.04, 0.07]</td>
</tr>
<tr>
<td>Noun test</td>
<td>-0.31</td>
<td>0.08</td>
<td>&lt; .001</td>
<td>[-0.48, -0.14]</td>
</tr>
<tr>
<td>IQ</td>
<td>0.73</td>
<td>0.19</td>
<td>&lt; .001</td>
<td>[0.35, 1.12]</td>
</tr>
<tr>
<td>NWR</td>
<td>0.23</td>
<td>0.14</td>
<td>0.09</td>
<td>[-0.04, 0.5]</td>
</tr>
</tbody>
</table>

**Significant aptitude x group interactions**

None

<table>
<thead>
<tr>
<th>PRODUCTION TASK – memory-based production of exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Main effects of aptitude measures</strong></td>
</tr>
<tr>
<td>LLAMA D</td>
</tr>
<tr>
<td>Noun test</td>
</tr>
<tr>
<td>IQ</td>
</tr>
<tr>
<td>NWR</td>
</tr>
</tbody>
</table>

**Significant aptitude x group interactions**

IQ × group 0.57 | 0.28 | 0.04 | [0.01, 1.14] | 1.77 |

<table>
<thead>
<tr>
<th>PRODUCTION TASK – pattern-based production of exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Main effects of aptitude measures</strong></td>
</tr>
<tr>
<td>LLAMA D</td>
</tr>
<tr>
<td>Noun test</td>
</tr>
<tr>
<td>IQ</td>
</tr>
<tr>
<td>NWR</td>
</tr>
</tbody>
</table>

**Significant aptitude x group interactions**

IQ × group -0.43 | 0.22 | 0.048 | [-0.86, 0.004] | 0.65 |

4.4 Discussion

The main goal of our study was to investigate if input rehearsal can foster learning of novel L2 patterns. The only study that investigated this question (Ellis & Sinclair, 1996) has suggested that this is the case, but due to design confounds, it remained unclear if benefits of rehearsal were due to the fact that learners had to utter the target structure, or because by uttering the structure, they created additional input for learning (Ellis & Sinclair, 1996).

Considering that this study investigated the role of rehearsal on pattern learning, it should first be pointed out that we found clear evidence of pattern learning only among learners who also developed awareness of the target
Effects of input rehearsal on L2 pattern learning

pattern. On the GJT that featured novel items, only aware learners performed above chance. On the GJT with familiar items, only aware learners clearly overgeneralized the pattern to exceptions, which is evidence of pattern knowledge (Tomasello, 2000). When it comes to the production task, there was potentially evidence of pattern learning without awareness because unaware learners – especially learners in the silent group – occasionally produced overgeneralized versions of exceptions. However, based on these results, it is not easy to draw solid conclusions. In any case, learners who were not aware of the pattern could not generalize the pattern to novel items, and they mostly seemed to have memorized determiner-noun combinations from the input, as the results of the GJT with familiar nouns suggest.

Based on our results, it is difficult to say why some learners acquired the pattern, while others did not. A possible explanation, also suggested by usage-based approaches (Abbot-Smith & Tomasello, 2006; Bybee, 2008) might be that learners’ accumulated knowledge of exemplars led to gradual pattern abstraction, and at some point to awareness. This would mean that not all learners in our study accumulated enough exemplar-based knowledge to be able to develop pattern knowledge and awareness. Also, different strategies that learners potentially adopted during the exposure may have had an additional effect. It is possible that some learners willfully adopted the memorizing strategy, while others adopted the strategy of pattern search during the exposure.

Performance of the aware learners in this study provided evidence that rehearsal led to better learning of determiner-noun combinations. However, based on our results, it is not entirely clear whether rehearsal fostered memorizing of determiner-noun combinations or also promoted learning and application of the underlying pattern. In both the production task and the GJT with novel nouns, we observed tendencies that point to the possibility that rehearsal helped aware learners to better apply the pattern they had noticed, and that it also diminished individual differences between learners. However, these results certainly need replication with larger sample sizes that would give us enough power to statistically detect the effects. The pattern that learners were exposed to in this study was unreliable, i.e., it featured exceptions, and exceptions have already been found to disturb pattern learning and the application of patterns, even when learners are aware of the pattern (as shown in Chapters 2 and 3). Rehearsal in this study made learners perform better on the GJT and production tasks, and it possibly attenuated the confusion caused by exceptions in the input. On top of this, results from the GJT with familiar nouns and the production task showed that rehearsal helped aware learners to memorize exceptions better, perhaps because aware learners were able to recognize exceptions as such. Our study provides evidence that the benefits of rehearsal are due to learners’ vocal repetition of the input. Given that the
rehearsal and silent group received identical amount of input, it must be the case that vocal rehearsal itself promoted learning.

The results of the current study point to a paradox. While learners in the rehearsal group generally seemed to learn more, they were less likely to become aware of the pattern, implying that rehearsing input may leave less room for becoming aware of the pattern. So, on the one hand, rehearsal seems to make noticing patterns more difficult – possibly because it takes learners’ attentional resources away from pattern learning, but on the other hand, once the pattern has been noticed, it leads to higher accuracies and possibly contributes to better and more consistent application of the pattern. Evidence that rehearsal promoted pattern learning and application, rather than just memorizing the input, may come from the fact that beneficial effects of rehearsal in this study were mostly observed in learners who were aware of the pattern but far less in learners unaware of the pattern. Thanks to rehearsal, aware learners seem to have been able to better apply the pattern to regular items and memorize exceptions on the top of that.

Our finding that rehearsal is not beneficial for the development of explicit rule awareness is not in line with the results of Ellis and Sinclair (1996). They found that learners in the rehearsal group had an advantage on a metalinguistic task. However, the important constraint of that study was that half of the learners in all groups received explicit rule explanation prior to instruction. Considering that not all learners in the study of Ellis and Sinclair received exposure without explicit rules, the results could not really tell us whether or not rehearsal fostered developing explicit metalinguistic knowledge.

It is not clear what the results of this study would mean for teaching. We showed that once aware, learners may benefit from rehearsal because they can better memorize regular items and exceptions, and also possibly better apply the pattern they had noticed. Given our results, a possibly favorable teaching approach could be to first encourage learners’ noticing of patterns, and then to use rehearsal activities to strengthen the learned pattern and acquire exceptions. To what extent it is practically feasible to implement rehearsal, and what the longer-term effects would be, remains to be seen. Also, similar effects may well be obtained if learners are asked to actively use the pattern in more meaningful production tasks, rather than to simply rehearse. In any case, it seems that rehearsal may be seen as one of the beneficial activities for learning, and may have a place in language instruction as an activity that supplements but does not replace more context-based activities (as pointed out in Hulstijn, 2001). Even though rehearsal may be a beneficial supplementary activity, it could be considered a boring one and therefore may be avoided. However, as shown in Hulstijn (2001), there are a number of ways in which rehearsal activities could be made more interesting.
The study also provides further support for the beneficial role of rehearsal in storing items in short-term memory (Baddeley, 2003, 2012). Namely, learners who were not aware of the target pattern and rehearsed input, showed some advantage in memorizing determiner-noun combinations compared to learners who did not rehearse. We also found clear evidence that different measures of memory capacity predicted learners’ performance on the GJT with familiar and novel items, and their production of the familiar items. This is in line with a number of studies that show that working memory is involved in language learning, and that rehearsal promotes memorizing vocabulary items by helping to keep the rehearsed input longer in short-term memory (Ellis & Beaton, 1993; Papagno et al., 1991; Seibert, 1927; Yoshida & Fukada, 2014). We also found some links between learners’ analytical abilities and learning of the target structure (both productively and receptively). For aptitude by treatment interactions, we observed some interesting interactions between IQ and group. Namely, in silent learners, higher IQ scores led to more pattern-based production of exceptions, whereas in the rehearsal learners, higher IQ scores led to more memory-based production of exceptions. This may imply that learners with better analytical ability could more easily recognize exceptions as such, memorize them better with the help of rehearsal, and in that way avoid overgeneralizations.

Several limitations of the study need to be taken into consideration. First, the learning session was relatively brief. Also, our sample of aware learners was rather small given that it was impossible for us to exercise control over how many learners would develop awareness. The rehearsal effects and their durability could be further investigated using longer exposure, delayed posttests, and larger samples that have enough power to give more solid evidence of the benefits of rehearsal for pattern learning. Second, in order to establish if learners were aware of the pattern, we used retrospective verbal reports. There has been some discussion about the validity of this approach because retrospective verbal reports might lead to an underestimation of the number of aware learners (Rebuschat et al., 2015). However, it should be pointed out that there is no perfect method of establishing awareness. For instance, as noted in Rebuschat et al. (2015), think-aloud protocols can also fail to detect aware learners, whereas trial-by-trial source attributions (Dienes & Scott, 2005) have been found to invite learners to search for the rules by suggesting rule as one of the options to choose from. For a detailed discussion on the advantages and disadvantages of different awareness measures, see Rebuschat et al. (2015).

To conclude, this study contributes to the small body of research on rehearsal, and makes some of the first steps in investigating the role of rehearsal on pattern learning. Further research will be needed to replicate the findings and extend them to other types of structures and to classroom settings.
However, the potential benefits or limitations of rehearsal activities should not be investigated by comparing instructions entirely based on drills with explicit and meaning-based kinds of instruction (e.g., Wong & VanPatten, 2003). Given that rehearsal activities should be used as the activities that are supplementary to context-based language instruction (Hulstijn, 2001), their effects need to be further investigated by looking at how they may or may not contribute to context-based types of instruction.
Chapter 5: The role of awareness, cognitive aptitudes, and instruction type in L2 processing of determiners

Abstract

This chapter investigates whether learners’ cognitive aptitudes (i.e., working memory and analytical ability), different levels of awareness, and instruction type (with or without rehearsal) can explain differences in second language processing of determiners, as measured by the visual world eye-tracking paradigm. Participants in the study received auditory exposure to a miniature language based on Fijian including a determiner-noun agreement pattern. Learners’ online processing of determiners was measured using a test with eye-tracking. Even though learners received only brief exposure to the language, we found evidence that some learners were able to anticipate the coming noun based on the determiner, and they also gained a speed advantage. We found that learners’ awareness played a crucial role in such anticipatory processing; only learners who were aware that determiners helped them during the test (i.e., prediction aware learners) showed signs of anticipatory processing. We found no evidence that either working memory or analytical ability modulated learners’ prediction ability, but we observed a tendency that working memory predicted whether learners would develop awareness of their own predictions. When it comes to the role of instruction type, we found evidence that exposure with rehearsal affected how determiners were processed. However, there was no clear evidence that it contributed to better processing.

5.1 Introduction

When investigating online language processing, researchers aim to explain how native speakers or second language learners process language as it unfolds in real time. This processing is usually measured in terms of ERP patterns, patterns of eye movements, or the pace at which speakers read or listen to language input. There has been a lot of research on processing abilities of second language (L2) learners (i.e., L2 processing) in order to see if learners can come to use certain L2 structures in similar ways as native speakers, and

*A slightly modified version of this chapter has been submitted as Curcic, M., Andringa, S., & Kuiken, F. (under review). The role of awareness and cognitive aptitudes in L2 processing of determiners.
under which conditions such nativelike processing may occur. Second language learners often show different language processing patterns from native speakers, and this is especially true for the processing of grammatical structures (Havik et al., 2009; Roberts, 2012). This means that it is difficult for adult L2 learners to learn to process grammar in ways similar to native speakers, but not impossible, as suggested by a number of studies (e.g., Dussias et al., 2013; Hopp, 2006; Osterhout et al., 2008).

Learners show substantial individual differences in how they process L2 structures (Roberts, 2012), and it is important to investigate which factors can explain these differences given that such knowledge can help us understand the outcomes of L2 learning in terms of processing. The factors that have been extensively investigated are L2 proficiency and L1 characteristics, and a relatively clear picture emerged showing that L2 proficiency and similarity between L1 and L2 structures influence learners’ L2 processing (e.g., Dussias et al., 2013; Dussias et al., 2014; Foucart & Frenck-Mestre, 2011; Sabourin & Stowe, 2008; Sagarra & Herschensohn, 2010; Tokowicz & McWhinney, 2005; Tokowicz & Warren, 2010). Also, there has been some evidence that working memory capacity contributes to grammar processing, i.e. that learners with higher working memory are more likely to show nativelike ways of processing (e.g., Dussias & Piñar, 2010; Havik et al., 2009; Williams, 2006). Other factors that could potentially explain individual differences in L2 processing, such as explicit knowledge or awareness and instruction types received very little attention in research so far.

The goal of this chapter is to investigate if several factors such as awareness, aptitudes, and instruction type can explain differences in beginner L2 learners’ online processing of determiners. There is evidence that these factors are related to success in language learning (see Leow & Donatelli, 2017 for a review on the role of awareness; Dörnyei, 2006 for the role of aptitudes; and De Graaff & Housen, 2009 for the role of instruction), but they have been rarely investigated in relation to online grammar processing. The two studies presented in this chapter (i.e., Study 1 and Study 2) investigate whether cognitive capacities or aptitudes (i.e., working memory and analytical ability) and learners’ awareness at different levels are predictive of determiner processing. Study 2 also looks at whether determiner processing can be improved through exposure in which learners are encouraged to rehearse language input.

5.1.1 Language processing in L1 speakers
Sentence processing in L1 speakers is claimed to be incremental, which means that sentences are processed in a word by word fashion, and listeners or readers make rapid online predictions about the upcoming linguistic material,
rather than wait until the end of a sentence to interpret it (Altman & Mirković, 2009). This makes language processing faster and more effective.

Evidence for incremental processing comes from a number of eye-tracking studies (see Huettig, Rommers, & Meyer, 2011 for a review) and self-paced reading/listening studies (e.g., Roberts & Felser, 2011; Sagarra & Herschensohn, 2010), in which listeners/readers have been shown to use certain linguistic elements to rapidly form predictions about what comes next in a sentence. Based on such predictions, they are able to quickly focus their eyes on the target object in the visual scene (in the visual world eye-tracking paradigm) or they are disturbed and slowed down by unexpected or implausible elements in a sentence (in the self-paced reading/listening paradigm). Such processing effects can be triggered by semantic features of words and pragmatics (e.g., Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003; Mani & Huettig, 2012). For example, Altmann and Kamide (1999) showed that when native speakers heard a sentence such as The boy will eat the cake, they were able to predict with their eyes that an edible object would follow, already before hearing the cake. Online incremental processing can also be triggered by grammatical features (e.g., Dussias, et al., 2013; Kamide, Scheepers, & Altmann, 2003; Lew-Williams & Fernald, 2007). In grammar processing research, determiners have been among the most frequently investigated grammatical structures, and determiner-based prediction has been found in several languages, such as Spanish (e.g., Dussias, et al., 2013; Lew-Williams & Fernald, 2007), French (e.g., Dahan, Swingley, Tanenhaus, & Magnuson, 2000; Foucart & Frenck-Mestre, 2011), and Dutch (e.g., Brouwer, Unsworth, & Mak, 2010; Loerts, Wieling, & Schmid, 2013; Sabourin & Stowe, 2008). For instance, Lew-Williams and Fernald (2007) found that both Spanish adults and very young children used grammatical gender information marked on articles to rapidly predict with their eyes the object that followed the article.

5.1.2 Language processing in L2 learners

In the last few decades, there has been an increasing interest in the online processing abilities of second language learners. A number of studies compared native speakers and L2 learners in order to investigate if L2 learners are able to process their L2 in nativelike ways, and what are similarities and differences between L2 and native processing. Accumulated evidence indicates that L2 speakers are sensitive to lexical and semantic information in similar ways as native speakers (e.g., Roberts & Felser, 2011; Williams, 2006; Williams, Möbius, & Kim, 2001). However, grammar is the area that seems to pose more difficulties to L2 processing, which means that L2 learners differ substantially from native speakers in how they process syntax and morphology, and they
also show substantial individual differences (Havik et al., 2009; Roberts, 2012). This observation led to the formulation of different hypotheses that tried to explain how L2 learners process grammatical structures, and whether they can ever reach nativelike levels of processing. For instance, the Shallow Structure Hypothesis (SSH) (Clahsen & Felser, 2006) proposed that late L2 language processing is shallow in that it mostly relies on lexical, semantic, and pragmatic cues, which is why L2 learners process grammatical structures in fundamentally different ways from native speakers. However, a number of studies provided evidence that L2 learners can show nativelike processing of grammar (e.g., Dussias et al., 2013; Hopp, 2006; Osterhout et al., 2008; Tokowicz & McWhinney, 2005; Tokowicz & Warren, 2010; Witzel, Witzel, & Nicole, 2009), and this led to the formulation of the Fundamental Identity Hypothesis (Hopp, 2007), according to which L2 learners can come parse and process grammatical structures in nativelike ways.

5.1.3 Factors contributing to L2 processing

Once accumulated research clearly showed that L2 learners sometimes can come to process L2 grammar in nativelike ways (e.g., Dussias et al., 2013; Hopp, 2006; Witzel et al., 2009), L2 processing research switched away from simply comparing native and non-native processing, and instead increasingly focused on investigating the factors that are involved in learners’ online processing, and factors that can potentially explain individual differences between learners.

As pointed out earlier, two factors received most attention so far: the level of proficiency in the L2 and the characteristics of the learners’ L1. These have been investigated using different methodologies, such as eye-tracking (e.g., Dussias et al., 2013), self-paced reading or listening (e.g., Jackson, 2008; Tokowicz & Warren, 2010), and ERPs (e.g., Dussias et al., 2014; Foucart & Frenck-Mestre, 2011; Sabourin & Stowe, 2008; Tokowicz & McWhinney, 2005), and among the most frequently used structures in this line of research have been gender-marked determiners. Accumulated research evidence has demonstrated that learners who have a high proficiency in L2 are more likely to show nativelike ways of processing (e.g., Dussias et al., 2013; Dussias et al., 2014; Jackson, 2008; Sagarra & Herschensohn, 2010), implying that online processing of grammatical structures is less likely to be found in beginner learners (e.g., Dussias et al., 2013; Dussias et al., 2014) although it is not impossible (e.g., Davidson & Indefrey, 2009; Tokowicz & Warren, 2010). In a similar way, learners are more likely to exhibit nativelike processing of L2 structures if their L1 has a similar structure that allows them to transfer L1 processing strategies to the L2 (e.g., Dussias et al., 2013; Foucart & Frenck-Mestre, 2011; Sabourin & Stowe, 2008; Tokowicz & McWhinney, 2005; Tokowicz & Warren, 2010). Dussias et al. (2013) investigated gender-related
incremental processing by comparing low and high proficiency L2 learners of Spanish to monolingual native speakers. The learners’ L1 was English, a language with no gender-marking on articles. The results showed that high-proficiency learners used gender information to predict the coming object in a way similar to native Spanish speakers, but the low-proficiency learners did not show such predictions. Additionally, the study included a group of low-proficiency learners of Spanish with L1 Italian (a language with gender-marking on articles). These learners showed gender-related anticipatory effects for feminine articles in spite of having low proficiency in Spanish, which implies that learners’ ability to make predictions is modulated by their L1 characteristics.

Individual differences in L2 processing can be attributed not only to learners’ proficiency levels and L1 characteristics, but potentially also to other factors, such as cognitive aptitudes. These other factors have been little investigated so far since researchers mostly tried to control for them, rather than look at how they influence processing (Roberts, 2012). Many studies have shown that factors such as awareness, cognitive aptitudes (e.g., working memory, analytical ability), and instruction type influence success in L2 learning (for reviews see Leow & Donatelli, 2017; Dörnyei, 2006; De Graaff & Housen, 2009). However, at present, there are few studies that investigated the relationship between these factors and learners’ online processing of particular L2 structures. Given that the two studies presented in this paper investigate the influence of aptitude components, awareness, and instruction type on L2 grammar processing, we will briefly review the existing research on the role of each of these three factors in L2 grammar processing.

When it comes to the role of aptitudes in online L2 processing, some research has been done on the role of working memory capacity. Working memory is known to contribute to better L1 comprehension and processing, as noted in the capacity theory of comprehension (Just & Carpenter, 1992; Just, Carpenter, & Keller, 1996). In L2 processing, as pointed out in Havik et al. (2009), little systematic research has been done, and very often no effects have been found (e.g., Felsner & Roberts, 2007; Juffs, 2004, 2005; Traxler, 2007). Still, some studies did find that learners with enhanced working memory capacity showed better L2 processing (Dussias & Piñar, 2010; Havik et al., 2009; Williams, 2006). For instance, Havik et al. (2009) used a self-paced reading task to investigate how German L2 learners of Dutch processed temporarily ambiguous subject and object relative clauses. They found nativelike processing patterns only among high working memory learners, and only on the task that required learners to make truth-value judgments about the argument roles of ambiguous nouns. Williams (2006) also found that only L2 learners of English who had high working memory capacity processed wh-questions in nativelike ways, as measured by a self-paced reading task. Similarly, in Dussias and Piñar
(2010), only learners with high working memory were able to use plausibility information in similar ways as native speakers to recover from the previous misparsing during their processing of long-distance wh-questions. These results suggest that working memory may be involved in L2 grammar processing, or at least L2 learning that leads to more nativelike processing. However, more research is needed before firm conclusions can be drawn, and also the role of other cognitive capacities needs to be addressed in the future research. For instance, analytical ability has often been found to contribute to the learning of grammatical patterns in L2 (e.g., Erlam, 2005; Harley & Hart, 1997; Robinson, 2005), but to our knowledge, its relation to L2 processing has not been investigated.

Another potentially important but rarely investigated factor that may lead to individual differences in L2 processing is learners’ awareness of the target structure or particular aspects of the structure. Andringa and Curcic (2015) investigated how learners’ awareness of a particular grammatical structure affected its processing, as measured by a task with eye-tracking. Learners in the study were briefly exposed to a miniature language based on Esperanto, and the target structure was differential object marking (DOM), according to which animate but not inanimate objects were preceded by the proposition a – “to”. Half of the learners were provided with a brief rule explanation during the exposure in order to raise their awareness of the target structure. The other half of the learners did not receive any rule explanation. The study found no evidence of nativelike predictive processing of DOM, and rule awareness only led to improved performance on a grammaticality judgment task, but not to predictive processing of DOM. Another study by Andringa and Curcic (in preparation) focused on highly proficient learners of Spanish, and investigated how learners’ awareness that DOM could be useful during the test affected their online processing. The test included two phases, and at the beginning of the second phase, learners were informed that DOM could help them during the test to arrive at the correct answer more quickly. Results provided clear evidence that learners, who hardly showed predictive processing in phase 1, clearly predicted in phase 2, after they had been made aware of the fact that DOM could be helpful to them. These two studies suggested that awareness of L2 structures during exposure may not directly translate to successful predictive processing. However, raising learners’ awareness that certain structures can be used during comprehension may directly translate to predictive processing of those structures, possibly because this level of awareness encourages learners to make strategic predictions. However, more systematic research is needed that looks at how awareness may or may not influence L2 processing.

When it comes to the question of how different instruction types may influence learners’ grammar processing abilities, there have been a few studies
that dealt with this question. Several studies have compared naturalistic and classroom L2 learners to investigate which instruction type is more favorable to L2 processing, and they found that immersion and naturalistic exposure to the L2 positively affected online syntactic processing (Frenck-Mestre, 2002; Pliatsikas & Marinis, 2013). Another line of research employed more controlled experimental designs to look at how different short instructional treatments may affect learners’ L2 processing. For instance, Marsden et al. (2013) investigated how explicit instruction with practice containing yes/no feedback affected learners’ online processing of morphology in a small artificial language. They found no evidence of morphological processing among learners, as measured by a cross-modal priming test. However, McManus and Marsden (2017) found that instruction consisting of explicit information and practice related to learners’ L1 (in addition to explicit information and practice related to their L2) led to improved processing of French imperfect, as measured by a self-paced reading task. Several studies used brief language exposure and ERPs to investigate how explicit vs. implicit instruction types affect learners’ L2 processing (Morgan-Short et al., 2010; Davidson & Indefrey, 2009; Batterink & Neville, 2013). Davidson and Indefrey (2009) showed that a short explicit instruction can lead to better L2 processing of adjective declension. However, this was not the case with the other structures investigated in their study: article-noun agreement and adjective-noun agreement in gender. Morgan-Short et al. (2010) and Batterink and Neville (2013) directly compared the effects of implicit and explicit instruction on L2 processing. Batterink and Neville (2013) found that the two instruction types did not differentially influence the processing of article-noun agreement, subject-verb agreement, and word order. Morgan-Short et al. (2010) found mixed results: the effects of implicit and explicit instruction on the online processing patterns of gender agreement structures depended on learners’ proficiency and type of structure. Knowing how different instructional treatments may influence processing of L2 structures can be important in understanding why learners differ in their L2 processing abilities, and how they come to develop them. Because of this, more research is needed that investigates L2 processing after different types of instruction.

In this chapter, we chose to investigate whether instruction with rehearsal of the input (i.e., overt repetition of stimuli) may lead to better L2 processing of gender-marked determiners. We hypothesized that rehearsing input may foster predictive online processing given that it helps to entrench input in working memory (Baddeley & Hitch, 1974), and working memory can foster online processing of grammar (e.g., Dussias & Piñar, 2010; Havik et al., 2009; Williams, 2006). Also, in Chapter 4, we showed that rehearsal fostered learning of determiner-noun combinations in the L2, so in this chapter we wanted to look
at whether rehearsal could also lead to better predictive processing of determiners.

5.1.4 The present studies
The two studies presented in this chapter are based on the data collected during the studies reported in Chapters 2, 3, and 4. In the present studies, we investigate if learners can develop predictive processing of determiners in their L2 after brief exposure to a novel language.

The main goal of Study 1 is to investigate how aptitudes (i.e., working memory and analytical ability) and awareness influence online L2 processing of gender-marked determiners. Study 2 replicates Study 1, and it also looks at the effects of instruction type, i.e. instruction with or without overt rehearsal of the input.

In both studies, participants were exposed to a novel miniature language based on Fijian and gender-based determiners within the language. In Study 2, we manipulated instruction type to see how rehearsing vs. not rehearsing input influenced predictive processing of determiners.

Learners’ processing of determiners was measured using the visual world paradigm with eye-tracking in order to check if learners could show predictive processing based on determiners. In the present studies, we focused on online processing abilities after a very brief language exposure. Learners’ L1, Dutch, also features a determiner system (de and het). Determiner de expresses common gender, whereas determiner het denotes neuter gender. This distinction allows for predictions, but there is no distinction between feminine and masculine gender, and there is no clear agreement between determiners and nouns (Booij, 2002). Given that the target structure was present in some form in learners’ L1, we expected predictive processing to be possible, even after short exposure.

In both studies, we administered two aptitude tests: one tapping into learners’ verbal working memory – LLAMA D (Meara, 2005) and the other tapping into analytical ability – non-verbal IQ test (Wechsler, 2008). Information on learners’ potential awareness of different aspects of the target structure was obtained through post experimental debriefing. We aimed to answer the following four research questions:

1) Can L2 learners develop predictive processing of determiners after brief exposure? (Study 1 and Study 2)

2) Can awareness of the target structure explain individual differences in learners’ predictive processing of determiners? (Study 1 and Study 2)
3) *Can working memory and analytical ability explain individual differences in learners’ predictive processing of determiners?* (Study 1 and Study 2)

4) *Can rehearsing input during instruction lead to higher rates of predictive processing of determiners?* (Study 2)

### 5.2 Method - Study 1

The dataset of Study 1 consisted of two data samples that were collected for the purposes of the studies presented in Chapters 2 and 3. These two data samples were collected at different moments in time, and the only difference between the two studies was the duration of exposure that learners received. For the purposes of this study, we decided to merge these two data samples in order to obtain more statistical power for answering our research questions. This was possible given that we found that the differences in exposure length in the two studies did not affect learners’ processing of determiners, as shown in an analysis that we present later.

#### 5.2.1 Participants

Participants in the study were 100 adult native speakers of Dutch, between 19 and 35 years old ($M_{\text{age}} = 23.59$, $SD = 3.76$). They were students or highly educated adults, but they were never trained in linguistics. All participants reported having good hearing, normal or corrected-to-normal vision, no history of dyslexia, and no color-blindness. The participants reported speaking at least one foreign language. In a questionnaire, we asked them to list the languages they spoke, and to assess the level of knowledge (A1 – C2 on the CEFR scales (Council of Europe, 2011)). These reports allowed us to estimate learners’ knowledge of foreign languages, and we computed two measures by awarding 0.5 points for each level of each language: 1) the FLK measure – reflecting the number and level of knowledge of all foreign languages, and 2) the AgrFLK measure – reflecting the number and level of foreign languages that featured gender-based determiners and agreement between determiner and noun. On average, learners scored 8.71 ($SD = 3.56$) on the FLK measure and 2.15 ($SD = 2.31$) on the AgrFLK measure.

At the end of the experiment, which lasted about 1 h and 45 min, learners were paid 15 euros for participation. The study was approved by the Ethics Committee of the University of Amsterdam.

#### 5.2.2 The target language and the target structure

The target language used in this study was a miniature language that we created by modifying Fijian – an Austronesian language of the Malayo-
Polynesian family, spoken in Fiji. The language consisted of 20 nouns, four verbs, six adjectives, and two determiners (Appendix A). All lexical items were based on Fijian, and where possible, both the original form and meaning were retained. However, whenever the original Fijian words had a complex form or reminded of words from participants’ L1 or potential L2s, we chose a form of another Fijian word.

The language had the same word order (i.e., SVO) as Dutch, and it included a determiner-noun agreement pattern. The present tense was marked by a –t suffix in the third person singular, and was created by adding the suffix to the Fijian verbs. The nouns were created by adding either –is or –uk to Fijian nouns. All nouns were preceded by one of the two definite determiners that were marked for gender: lep – a masculine determiner or ris – a feminine determiner. The determiner agreed with the noun in that lep preceded nouns ending in –uk (e.g., lep oseuk (“the ball”)), whereas ris preceded nouns ending in –is (e.g., ris salis (“the chair”)). The participants’ L1 – Dutch – also has two definite determiners (de – common and het – neuter), but there is no feminine vs. masculine distinction between them, and also no clear determiner-noun agreement (Booij, 2002). Dutch native speakers have been shown to predict the coming objects based on determiners in their L1 processing (Brouwer et al., 2010; Loerts et al., 2013; Sabourin & Stowe, 2008).

5.2.3 Language exposure
Participants were told that they would be learning a new language by looking at images and listening to sentences describing the images. They were encouraged to learn the language as well as they could because their knowledge would be tested from time to time. Learners were not informed about the purpose of the experiment or the existence of the target structure. The exposure started with a noun learning phase, followed by a noun test. After this, learners were exposed to phrases and sentences containing the target structure.

We used E-prime software and a Tobii TX120 computer to present the materials. All instructions that learners received in the course of the experiment were written on the computer screen in Dutch, and were also recorded by a female native speaker of Dutch. The experimental materials were always presented auditorily, and were recorded by a female native speaker of Serbian. The images used in the experiment were retrieved from the Clipart image database (Clipart, Vector Graphics, and Illustrations, 2014) and edited in Photoshop.

Noun learning and assessment
First, we wanted to teach learners the nouns of the new language and make sure that all learners had identical and perfect knowledge of nouns before
receiving exposure to the determiner-noun combinations. In the noun learning phase, learners were exposed to 20 bare nouns. In every trial, they saw a simple, black-and-white image, and they heard the noun denoting the object in the image. Every noun appeared six times, and the order of the trials was identical for all learners.

After the noun learning phase, we tested learners’ knowledge of the nouns. In each trial of the test, learners saw four images on the screen and heard a noun. They were asked to click on the correct image that corresponded to the noun they had heard, after which they were given feedback about the accuracy of their response so that they could continue learning. Each noun appeared as the target noun in two trials, whereas the distractor images were randomly chosen by the program. The first 40 trials were identical for all learners, after which they received additional trials with nouns that they had not recognized correctly. The test continued until learners reached 100% accuracy. We used learners’ scores on the noun test as a measure of their rote memory ability. However, considering that the minimum number of items on this test was 40, the test did not measure all the potential differences in rote memory ability for learners who scored 40. Results of the noun test are presented together with the other aptitude measures in section 5.2.6.

Target structure exposure
In this part of the exposure, learners received more complex linguistic input. Next to the 20 nouns taught in the previous phase, it included two determiners, three intransitive verbs, copula be, and six adjectives (Appendix B). In every trial, learners saw an image representing a simple object, an intransitive activity, or a state, and they heard a phrase or sentence describing the image. The trials fell into one of the following four structural categories:

1) Determiner + noun

![Image of a chair]

*Ris salis*
‘The chair’
2) Determiner + adjective + noun

Ris matene salis
"The blue chair"

3) Intransitive sentences: determiner + animate noun + verb

Ris burogis sisilit.
"The dog is swimming."

4) Intransitive sentences: determiner + inanimate noun + is + adjective

Ris salis na duka.
"The chair is broken."

Learners heard 308 trials, and the exposure was divided into two parts with a break in the middle. The first part consisted of 176 trials, and lasted 15 minutes. The second part lasted 10 minutes, and learners saw 132 trials. Eighteen out of 20 nouns appeared 14 times, and two nouns were overrepresented and appeared 28 times each: feminine noun burogis – “dog”, and masculine noun dawauk – “cat”. The reason why two nouns were overrepresented was to make the input resemble natural input more (Ellis & Ferreira-Junior, 2009). The order of the trials was fixed and identical for all learners.

As mentioned earlier, half of the learners received longer exposure in that their instruction included an additional set of 176 trials. These were the 176 trials from the first part of the exposure, but with a different ordering.
For the purposes of the studies presented in Chapters 2 and 3, the present study included an additional variable – reliability of items. Although this variable is of no interest to the present study, it is important to explain it here and later check for its effects. For half of the learners \( (n = 50) \), 4 out of 20 nouns – two feminine and two masculine – were exceptions to the agreement pattern between determiners and nouns. For instance, the noun touk was presented with the masculine determiner lep to the reliable input learners, whereas it was presented with the feminine determiner ris to the unreliable input learners.

Learners received equal exposure to both feminine and masculine items, and the two determiner categories had the same number of overrepresented vs. underrepresented nouns, animate vs. inanimate nouns, non-biological vs. biological gender nouns (i.e., boy, girl, man, woman).

### 5.2.4 The test with eye-tracking

The goal of the test with eye-tracking was to see whether learners made rapid eye predictions about the coming noun based on the determiner they heard. We also wanted to see if learners had a speed advantage in processing linguistic materials when determiners were predictive of the coming noun. For this purpose we recorded learners’ reaction times (RTs).

The test materials consisted of 40 trials (Appendix E). In each trial, learners saw two images on the screen – one on the left and one on the right side, and they heard a short phrase in the new language, referring to one of the two images. The phrase consisted of a determiner, an adjective, and a noun (e.g., lep ramase karetuk – “the green bird”). The learners’ task was to choose the correct image by pressing either “z” (for the left image) or “m” (for the right image) on the keyboard. They were encouraged to be as fast and accurate as possible. All images were familiar to learners from the previous exposure phase.

In the construction of the test trials, we systematically varied gender of the target (correct) noun - masculine vs. feminine, and the gender of the images in the visual scene - same gender vs. different gender (Appendix E). Every trial fell into one of the following four experimental conditions:
1) Masculine same-gender trials

*Lep ramase oseuk*
“The green ball”

Target image (masc.)  Distractor image (masc.)

2) Masculine different-gender trials

*Lep matene kumiuk*
“The blue window”

Distractor image (fem.)  Target image (masc.)

3) Feminine same-gender trials

*Ris galile sitois*
“The red rabbit”

Target image (fem.)  Distractor image (fem.)
4) Feminine different-gender trials

*Ris matene ikais*
“The blue computer”

There were 10 trials per condition. In every condition, there were eight trials featuring the regular nouns, and two trials featuring the nouns that were exceptions for the unreliable input group. Given that these trials differed for the reliable and unreliable input learners, they were later excluded from the analysis for both learner groups. Every noun appeared twice as the target noun – once in the same gender condition, and once in the different gender condition. Four different color adjectives that were previously introduced in the exposure phase were used in the test, and they appeared equally often. Also, the target images were presented on the left and on the right side of the screen equally often.

While constructing the test materials, we took care not to bias participants’ eye-gazes and responses towards one of the two images in the visual scene. Given that in visual processing animate objects may have advantage over inanimate objects (e.g., Proverbio, Del Zotto, & Zani, 2007; Thomas & Forde, 2006), we never combined images of animate and inanimate objects in the visual scene. Also, the images of the four nouns that had biological gender (man, woman, girl, and boy) were always combined with the images of other biological gender nouns in the visual scene. The reason was that biological gender nouns have more cues to their gender class than nouns without biological gender. Within a single trial, we always used the same color for target and distractor images, meaning that the color adjective was never indicative of the correct answer, and learners could predict the correct image only based on the determiner.

Before the test started, the eye-tracker was calibrated to participants’ eyes, and it recorded participants’ eye movements at a rate of 120 Hz. After this, participants did four practice items to make sure they understood the task. At the beginning of each trial in the test, participants saw a fixation cross in the middle of the screen, and after they had fixated their eyes on it for 100 ms, the trial started. In this way we made sure that participants’ eyes were in the middle of the screen before each trial started. In each trial, images were
presented for about 1200 ms before the sound was played, and learners had a maximum of 5000 ms to choose the correct image.

If participants had learned to use determiners to predict the upcoming information, we expected that they would focus their eyes on the target image more than on the distractor image already while hearing the determiner and adjective, and before hearing the target noun. We expected this to occur in different-gender trials only, where the determiner was predictive of the correct answer, but not in the same-gender trials. During the noun phase, learners were expected to focus on the correct image. In terms of RTs, if learners obtained a speed advantage from determiner processing, we expected that they would choose the correct image faster – and thus have shorter RTs – in different-gender trials than in the same-gender trials.

Learners had an average accuracy of 96.78 % (SD = 4.11, min = 83.94, max = 100) when choosing the target image, which means that they were able to recognize the target nouns very accurately.

5.2.5 Debriefing
After the language exposure and tests, we debriefed learners using a protocol in order to gauge their awareness of different aspects of the target structure. We hypothesized that three observed levels of awareness could potentially lead to better predictive processing of determiners: awareness of the gender distinction in the language (i.e., that *lep* is a masculine determiner, and *ris* is a feminine determiner), awareness of the agreement pattern between determiners and nouns, and awareness that determiners were useful during the test. At the end of debriefing, we explained the purpose of the experiment, and we told learners that the language was a non-existing language based on Fijian.

5.2.6 Aptitude measures
Three aptitude measures were administered to the entire sample: a measure of analytical ability (IQ test), working memory (LLAMA D), and rote memory (noun test). Our goal was to check if these aptitudes could predict learners’ ability to develop predictive processing of determiners. We will explain the aptitude measures in the order in which they were administered.

*The noun test*
On average, learners needed 44.89 items (SD = 5.91, min = 40, max = 71) to pass the test, and on this test lower scores indicated better performance. The test scores gave some indication of learners’ rote memory ability, but since 21 out of 100 learners scored 40, the test did not measure the full range of differences in learners’ rote memory ability.
The role of awareness, aptitudes, and instruction type in L2 processing

The non-verbal IQ test
The Matrix Reasoning subtest of the Wechsler Adult Intelligence Scale – 4th edition (WAIS-IV) (Wechsler, 2008) was used as a measure of learners’ analytical ability. The test measures non-verbal abstract problem solving and inductive reasoning ability. The test consisted of 26 items, and the learners’ task was to choose the missing part of a visual pattern. There were five options to choose from. The test was administered after the first part of the exposure to the target structure, and participants were given maximum 20 minutes to complete the test. On average, learners scored 22.99 ($SD = 1.86$, min = 17, max = 26) on this test.

The LLAMA D sound recognition task
The LLAMA D task was used to assess learners’ working memory. The test is a part of the LLAMA language aptitude test battery (Meara, 2005). It measures the ability to recognize spoken language that you have been exposed to before. Participants first listened to a sequence of 10 words in an unknown language, after which they did a 30-item test in which they heard both previously heard words and novel words. Their task was to decide for each word whether they had heard it before. If they decided correctly, they gained points, and if they decided wrongly, they lost points. The maximum score was 75, and the task lasted about 5 minutes. It was administered after the debriefing session. On average, learners scored 27.85 on this test ($SD = 13.75$, min = 0, max = 65).

5.3 Results – Study 1
The goal of Study 1 was to investigate whether learners can develop predictive processing of determiners after a brief exposure, and if so, whether different aptitudes and levels of awareness can explain individual differences in this processing. We start this section by presenting the results of debriefing, which helped us construct awareness-based subgroups of learners. We then present the online processing measures (eye movements and RTs), and how learners’ predictive processing was affected by different awareness categories and different aptitudes.

5.3.1 Debriefing results – awareness
Most learners reported that they focused on learning the nouns, verbs, and adjectives during the exposure, and paid less attention to the other aspects of the language. As Figure 5.1 shows, more than half of the learners – 62 out of 100 – reported that they had become aware of gender distinction during the exposure, i.e. that lep was a masculine determiner, whereas ris was a feminine determiner. For the sake of presentation, learners with this level of awareness
will be referred to as *gender aware learners*. Twenty-four learners noticed the agreement pattern between the determiner and the noun during the exposure, and were able to reproduce it during debriefing; these learners are labeled *pattern aware learners*. As can be seen from Figure 5.1, it was mostly the case that pattern aware learners were also gender aware. There was only 1 learner who noticed the pattern, but was unaware of the gender distinction. Some learners also reported being aware that determiners sometimes helped them during the test to arrive at the correct answer faster \((n = 16)\. We refer to these learners as *prediction aware learners*. These learners were always also gender aware, and were in some cases \((n = 10)\. but not always \((n = 6)\. also aware of the pattern, as shown in Figure 5.1. The results of debriefing imply that gender awareness was a prerequisite for developing the other, deeper levels of awareness in our study.

**Figure 5.1.** Overview of learners’ different levels of awareness, as reported during debriefing (Study 1)

### 5.3.2 Eye movement results

During the test with eye-tracking, Tobii recorded the position of participants’ eyes every 8.3 ms, and for each 8.3 ms frame, we marked gaze accuracy as 1 – if a participant looked at the target image, or 0 – if a participant looked at the distractor image. Frames for which eye data were missing were marked as *missing*, whereas the frames in which participants’ eyes were focused neither on the target nor distractor image, were marked as *irrelevant*.

Altogether 26.6% of the frames were missing or irrelevant \((17.84\%\text{ missing}, 8.76\%\text{ irrelevant})\. and these were not included in the analyses. Before analyzing the data, we excluded all trials in which more than 50% of the frames were missing or irrelevant. Using this procedure, 17.38% of the trials were
excluded. We then excluded the participants that lost more than two thirds of all test trials (i.e., 70%). In this way 4 participants were excluded.

The goal of the statistical analyses was to see if learners could use determiners to anticipate the correct answer with their eyes while hearing the determiner and adjective, but before hearing the disambiguating noun. We expected learners to show a preference for the target image only in the different-gender trials; we expected no preference in the same-gender trials because the determiner was not predictive of the correct image. The determiner-adjective and noun phases were equally long in the same and different-gender trial sets, because each determiner + adjective combination and each noun occurred equally often in the two trial sets. Also, the same recordings for each determiner + adjective combination and for each noun were used in the two trial sets. The average boundary between the determiner-adjective phase and the noun phase was at about 980 ms after the onset of the determiner (SD = 79). We applied a correction of 200 ms to the phases in order to account for the standardly observed delay in adult eye-responses to linguistic materials (Dahan, Magnuson, Tanenhaus, & Hogan, 2001). The figures that we present in the results section already include the 200 ms correction. Also, whenever we refer to time windows in which predictive processing is observed, the references already include the 200 ms correction.

To analyze the gaze data, we used the cluster-based permutation analysis (described in Dink & Ferguson, 2015; used in Maris & Oostenveld, 2007) from the eyetrackingR package (Dink & Ferguson, 2015). The advantage of this analysis is that it allowed us to establish the onset and duration of predictive processing. This statistical procedure involves two main steps. In step 1, statistical tests are run on time bins of a particular size in order to quantify the significance of an effect at each time bin. All adjacent bins where the effect is significant are grouped into time clusters, and for each time cluster, the sum of the statistics for the time bins inside the cluster is calculated. If time clusters with significant effects are found in step 1, then step 2 of the analysis is performed. Here data are randomly shuffled, and the testing procedure is repeated. This shuffling and testing is repeated hundreds of times, which eventually gives an indication of what clusters could be expected if there was no effect (i.e., randomly shuffled data). These results are finally compared to the results from step 1 to obtain a p-value for each initially detected time cluster, which shows whether the initially detected time clusters contain significant effects or not. This procedure allows saving data sensitivity, while applying good corrections to avoid false alarms. In our analyses, we used 50 ms time bins, and the procedure of reshuffling and testing was repeated 2500 times.

In the analysis, we first looked at whether overall predictive processing effects were observed. For this, we analyzed the 50 ms time bins using paired samples t-tests, where we compared the proportion of looks towards the target
image in the same-vs. different-gender trials. Then we went on to look at whether predictive processing was modulated by different levels of awareness. The cluster-based permutation analysis does not allow for investigating interaction effects, but it was possible to run these analyses by creating a dependent variable that we called prediction score, and that captured learners’ predictive processing. The prediction scores were obtained by subtracting the proportion of looks towards the target image in the same-gender trials from the proportion of looks in the different-gender trials, in each 50 ms time bin, for each participant. We then analyzed time bins using linear models to assess the effect of awareness on the prediction scores, i.e. whether awareness influenced learners’ predictive processing. Finally, we analyzed the role of the aptitudes in predictive processing. For this purpose, we again used prediction scores as a dependent variable, and for every time bin, we used a linear model to see whether aptitude scores predicted the prediction scores.

Before analyzing the results of the test with eye-tracking, we excluded all exception test trials from both the reliable input and the unreliable input learner groups, since there was a possibility that these trials were processed differently by the unreliable input group. As a first step in the analysis, we checked whether there was an effect of exposure length on prediction scores. We found that the length of exposure did not affect learners’ predictive processing of determiners, which means that we could merge the two datasets for the purpose of further analysis. Also, the type of input learners had received (reliable vs. unreliable) had no significant influence on predictive processing, so we merged these two groups of learners in further analyses.

Next, we analyzed the gaze data to see if learners as a group showed predictive processing of determiners. The dependent variable was the proportion of looks towards the target image, and the within-subject factor was trial type (same vs. different gender). In step 1 of the analysis, we found a time cluster indicating the presence of predictive processing between 550 and 1250 ms after the onset of the crucial determiner-adjective phase. In step 2 of the analysis, this effect was confirmed as significant (Figure 5.2; Table 5.1). The two black vertical lines in Figure 5.2 show phase boundaries after applying the correction of 200 ms. This means that learners as a group showed predictive processing of determiners about 400 ms before they could reasonably start responding to the nouns. They looked more at the target than the non-target images in the different-gender trials, as compared to the same-gender trials, and this effect spilled over into the noun phase (Figure 5.2).
The role of awareness, aptitudes, and instruction type in L2 processing

Figure 5.2. Proportion of learners' looks towards the target image over time in the same-gender and different-gender trials, and the predictive processing effect observed between 550 and 1250 ms (shaded area)

Table 5.1. Results of the cluster-based permutation analysis – overall prediction effects

<table>
<thead>
<tr>
<th>Time cluster</th>
<th>Sum statistic</th>
<th>Time range in ms</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31.72</td>
<td>550 – 1250</td>
<td>0.003</td>
</tr>
</tbody>
</table>

As a next step, we wanted to see if learners' ability to process determiners online was modulated by awareness. Using the debriefing results as the starting point, we looked at the effects of a) gender awareness, b) pattern awareness, and c) awareness of using determiners during the test (i.e., prediction awareness). In each analysis, one of the three levels of awareness was entered as a fixed effect. The dependent variable was always the prediction score that expressed the difference in the proportion of looks to the same- and different-gender trials. We found no effects of gender awareness and pattern awareness. In step 1, the analyses identified no significant initial time clusters that would point to potential effects of these two levels of awareness. However, we did find a significant effect of prediction awareness, such that learners who had reported this awareness \((n = 16)\) showed higher levels of predictive processing than those who had not \((n = 80)\). In step 1 of the analysis, several potential time clusters were detected, but only the largest one turned out to be significant.
after step 2 (Table 5.2). This effect occurred between 350 and 850 ms after the onset of the determiner-adjective phase, and is indicated by the shaded area in Figure 5.3. It means that learners who were prediction aware showed earlier and stronger predictive processing compared to learners without this level of awareness (Figure 5.3; Table 5.2).

**Figure 5.3.** Proportion of learners’ looks towards the target image over time in the same-gender and different-gender trials, split by learners’ prediction awareness, and the time cluster during which the effect of prediction awareness was observed (shaded area)

**Table 5.2.** Results of the cluster-based permutation analysis – effects of prediction awareness

<table>
<thead>
<tr>
<th>Time cluster</th>
<th>Sum statistic</th>
<th>Time range in ms</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>250 – 300</td>
<td>0.53</td>
</tr>
<tr>
<td>2</td>
<td>24.86</td>
<td>350 – 850</td>
<td>0.007</td>
</tr>
<tr>
<td>3</td>
<td>2.03</td>
<td>900 – 950</td>
<td>0.52</td>
</tr>
<tr>
<td>4</td>
<td>2.06</td>
<td>1150 – 1200</td>
<td>0.49</td>
</tr>
</tbody>
</table>

In order to check if there was any predictive processing of determiners by learners who were unaware that determiners were helpful during the test, we
ran a separate analysis on learners without prediction awareness. Step 1 of the analysis detected two potentially significant time clusters: between 850 and 900 ms, and between 1050 and 1100 ms. However, step 2 showed that these effects were not significant, meaning that we can treat the detected clusters from step 1 as chance findings (Table 5.3). This implies that we found no evidence of predictive processing without awareness.

Table 5.3. Results of the cluster-based permutation analysis – online processing effects of prediction unaware learners

<table>
<thead>
<tr>
<th>Time cluster</th>
<th>Sum statistic</th>
<th>Time range in ms</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.13</td>
<td>850 – 900</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>2.01</td>
<td>1050 – 1100</td>
<td>0.53</td>
</tr>
</tbody>
</table>

When it comes to the relation between learners’ predictive processing and the aptitude measures, we found no effects of working memory, rote memory, and analytical ability. In these analyses, step 1 revealed some potentially significant time clusters, during which aptitude scores potentially modulated the prediction score. However, in step 2, these effects turned out not to be significant, as there was between 30 and 50% probability that they had been detected by chance in step 1 (Table 5.4).

Table 5.4. Results of the cluster-based permutation analysis – effects of aptitude measures on predictive processing

<table>
<thead>
<tr>
<th>Time cluster</th>
<th>Sum statistic</th>
<th>Time range in ms</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLAMA D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.14</td>
<td>550 – 600</td>
<td>0.44</td>
</tr>
<tr>
<td>2</td>
<td>2.1</td>
<td>1700 – 1750</td>
<td>0.46</td>
</tr>
<tr>
<td>Noun test scores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.08</td>
<td>1250 – 1350</td>
<td>0.33</td>
</tr>
<tr>
<td>IQ – analytical ability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.39</td>
<td>600 – 700</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Finally, we wanted to check if the aptitude scores predicted learners’ awareness at different levels, and we used generalized linear models (GLMs) to do so. We found that learners’ gender awareness was predicted by rote memory, as measured by the noun test ($b = -0.1, SE = 0.04, p = .02, 95\% CI [-0.19, -0.02]).
0.18, -0.02], \textit{OR} = 0.91), and also by analytical ability, as measured by the IQ test (\( b = 0.24, SE = 0.12, p = .04, 95\% \text{ CI} \ [0.003, 0.47], \textit{OR} = 1.27 \)). Learners’ pattern awareness was predicted by the LLAMA D measure of working memory (\( b = 0.06, SE = 0.02, p = .002, 95\% \text{ CI} \ [0.02, 0.1], \textit{OR} = 1.06 \)), and there was a tendency that learners’ prediction awareness was predicted by the LLAMA D scores (\( b = 0.04, SE = 0.02, p = .066, 95\% \text{ CI} \ [-0.003, 0.08], \textit{OR} = 1.04 \)). The tendency that learners with higher LLAMA D scores were more likely to develop prediction awareness may imply that working memory may have had an indirect positive influence on predictive processing via awareness. It is possible that this tendency did not reach significance due to the small number of learners who were prediction aware – 16 learners out of 100.

### 5.3.3 Reaction times (RTs)

The RT analysis was conducted on the correct responses only, and again exception items were excluded. The goal was to find out whether learners gained a speed advantage in choosing the correct image when determiners were predictive of the correct response, and how different levels of awareness and aptitudes modulated the speed gains. The RTs were measured from the onset of the noun until the button press. Since the same-gender and the different-gender trials featured identical nouns, the comparisons of RTs across the two trial sets gave objective differences in RTs.

To analyze the RTs, we used linear mixed models, which included both random and fixed effects. The dependent variable was response time (RT) in milliseconds. The fixed effect was always trial type (same gender vs. different gender). The other fixed effects were the awareness and aptitude variables because we wanted to check whether these variables modulated the speed advantage. Following the recommendation of Barr et al. (2013), we always applied maximal random effects structures. This means that we always included both by-subject and by-item random intercepts and slopes for each fixed effect included in the model. Whenever subjects or items were nested within a particular fixed effect variable, we did not include the corresponding by-subject or by-item random slopes because such models would be unidentifiable. For instance, given that trials were nested within trial type (same vs. different gender), we included by-subject but not by-item random slope for the trial type fixed effect.

RT results largely confirmed the eye-tracking results. We found that there was no main effect of trial type – learners as a group did not gain speed advantage when determiners were predictive of the correct answer (\( b = -53.74, SE = 48.14, t = -1.12, p = .26, 95\% \text{ CI} \ [-148.61, 41.1] \)).

When it comes to the effects of different levels of awareness on speed gains, we found that gender awareness did not lead to a speed advantage, i.e.
there was no significant interaction between trial type and gender awareness ($b = -44.67, SE = 47.08, t = -0.95, p = .34, 95\% CI [-138.84, 49.49]$). However, both pattern awareness and prediction awareness led to a speed advantage: there were significant interactions between trial type (same vs. different gender) and pattern awareness ($b = -132.25, SE = 62.16, t = -2.13, p = .03, 95\% CI [-256.58, -7.92]$); and between trial type and prediction awareness ($b = -274.75, SE = 66.3, t = -4.14, p < 0.001, 95\% CI [-407.35, -142.14]$). Given that pattern awareness and prediction awareness overlapped substantially, it was possible that the effect of pattern awareness was driven by the learners who were also prediction aware. Therefore, we checked if pattern aware, but prediction unaware learners ($n = 14$) gained a speed advantage. We found no speed advantage in this subgroup, i.e. no main effect of trial type ($b = -1.02, SE = 62.98, t = -0.02, p = .99, 95\% CI [-126.99, 124.95]$), which indicates that only prediction awareness really contributed to learners’ speed gains. As Table 5.5 shows, when learners were prediction aware, they were able to choose the correct image almost 300 ms faster in the different-gender trials than in the same-gender trials.

<table>
<thead>
<tr>
<th>Prediction awareness</th>
<th>Different-gender trials</th>
<th>Same-gender trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aware ($n = 16$)</td>
<td>1869 (498)</td>
<td>2152 (216)</td>
</tr>
<tr>
<td>Unaware ($n = 84$)</td>
<td>2294 (237)</td>
<td>2303 (259)</td>
</tr>
</tbody>
</table>

For the effects of the aptitudes on speed advantage gains, we found no evidence that any of the aptitude measures modulated learners’ speed gains, i.e. there were no interactions between trial type and LLAMA D scores ($b = 0.32, SE = 1.65, t = 0.19, p = 0.85, 95\% CI [-2.98, 3.61]$), trial type and noun test scores ($b = -0.95, SE = 4.16, t = -0.23, p = .82, 95\% CI [-9.27, 7.37]$), and trial type and IQ scores ($b = -14.5, SE = 12.04, t = -1.2, p = 0.23, 95\% CI [-38.59, 9.58]$). However, we observed a main effect of the noun test, i.e. noun test scores predicted the speed of performance on all items ($b = 13.39, SE = 5.84, t = 2.29, p = .02, 95\% CI [1.72, 25.07]$). This effect implies that learners who needed more items to pass the noun test, later also needed more time to recognize the nouns during the test with eye-tracking.
5.4 Method – Study 2

5.4.1 Participants
Participants in the study were 50 adult native speakers of Dutch, between 19 and 35 years old ($M_{age} = 23.14, SD = 2.9$). Like in the Study 1, participants were students or highly educated adults, without training in linguistics. All participants reported having good hearing, normal or corrected-to-normal vision, no history of dyslexia, and no color-blindness. We randomly assigned participants to one of the two instruction groups: the rehearsal group (7 males, 18 females, $M_{age} = 23.76, SD = 3.5$) and the silent group (6 males, 19 females, $M_{age} = 22.52, SD = 2.06$). Learners in the rehearsal group were asked to repeat every sentence or phrase they heard aloud, whereas learners in the silent group heard every sentence or phrase twice, without repeating it.

Learners reported speaking at least one foreign language. According to the Welch two-sample t-test, the rehearsal ($M = 8.58, SD = 3.29$) and the silent learner groups ($M = 9.86, SD = 3.37$) did not differ significantly in their knowledge of all foreign languages – the FLK measure ($t(47.97) = 1.36, p = .18, 95\% \ CI \ [-0.61, 3.17]$). When it comes to the knowledge of languages with determiner-noun agreement – AgrFLK measure, a non-parametric Mann-Whitney U test showed that the rehearsal ($M = 2.42, SD = 1.99$) and the silent group learners ($M = 2.52, SD = 2.17$) also did not differ significantly ($W = 319.5, p = .90, 95\% \ CI \ [-1, 1]$).

The experiment lasted about 1 h and 45 min, and learners were paid 15 euros for participation. The study was approved by the Ethics Committee of the University of Amsterdam.

5.4.2 Language exposure
The target language and structure used in this study were identical to the ones used in Study 1. Again, learners were first exposed to nouns, and this phase was followed by a noun test. The purpose of the noun test was to make sure that all participants developed perfect knowledge of the nouns, and also to measure the participants’ rote memory. Finally, learners were exposed to short phrases and sentences containing the target structure.

Language exposure was nearly identical to the one in Study 1, but there were a few minor differences. During the noun learning phase, learners in both groups were asked to repeat aloud every noun they heard. During the exposure to the target structure, learners heard 308 trials with a short break in the middle. Learners from the rehearsal group heard each phrase and short sentence once, and were given a few seconds to repeat it aloud, after which the

---

4 Whenever t-test assumptions (i.e., normality of distribution and homogeneity of variance) were violated, we used a non-parametric Mann-Whitney U test instead of a t-test.
next trial started. Learners from the silent group heard each phrase or sentence twice. For the rest, the exposure was identical to the one in Study 1. The instruction lasted equally long for the rehearsal and silent groups of learners, and both groups received an identical amount of input if we assume that the output of the rehearsal group also served as input. Thus, the two groups differed only in the presence or absence of overt rehearsal. During the exposure, all learners from the rehearsal group repeated phrases and sentences faithfully.

For the purposes of another study, we made input slightly irregular for both groups of learners, which means that 4 out of 20 nouns – two feminine and two masculine – were exceptions to the agreement pattern between determiners and nouns, i.e. in exception items, -is nouns were preceded by lep, and -uk nouns were preceded by ris.

5.4.3 The test with eye-tracking
We used the same eye-tracking materials and procedures as in Study 1. Learners had an average accuracy of 97.04% (SD = 4.64, min = 77.27, max = 100) when choosing the correct image, which means that they could recognize the target nouns very accurately. According to a Mann-Whitney U test, the rehearsal (M = 96.33, SD = 5.42) and the silent group (M = 97.75, SD = 3.67) did not differ significantly in their performance (W = 354, p = .39, 95% CI [-0.0004, 0.03]).

5.4.4 Debriefing
After the language exposure and tests, we used the same debriefing protocol as in Study 1 to check if learners were aware of different aspects of the target structure.

5.4.5 Aptitude measures
We administered the same measures of learners’ working memory (LLAMA D), rote memory (noun test), and analytical ability (IQ test) as in Study 1.

On average, learners scored 28.1 on LLAMA D (SD = 12.73, min = 0, max = 55). A Welch two-sample t-test showed that there was no significant difference between the rehearsal (M = 30, SD = 13.46) and the silent group (M = 26.2, SD = 11.93) in their performance on LLAMA D (t(47.31) = -1.06, p = .30, 95% CI [-11.04, 3.44]).

Learners needed on average 45.26 items (SD = 6.63, min = 40, max = 76) to pass the noun test, and according to a Mann-Whitney U test, learners in the rehearsal (M = 46.32, SD = 5.94) and the silent group (M = 44.2, SD = 7.22) did not differ significantly (W = 218.5, p = .07, 95% CI [-4, 0]).
5.5 Results – Study 2

The goal of Study 2 was to investigate whether different instruction types (input rehearsal vs. absence of rehearsal) can influence learners' ability to process determiners online. In addition, in this study, we again looked at whether aptitudes and different levels of awareness could explain individual differences in learners' ability to process determiners. The awareness-based subgroups of learners were created based on learners' reports from debriefing. We will first present the debriefing results, and then look at the eye movement results and RTs.

5.5.1 Debriefing results – awareness

As in Study 1, learners mostly reported that they focused on learning nouns, verbs, and adjectives during the exposure. As Figure 5.4 shows, most learners – 37 out of 50 – developed gender awareness during the exposure. Fourteen learners also noticed the agreement pattern between the determiner and the noun, and all learners who were pattern aware were also gender aware. There were 9 learners who reported the awareness that determiners sometimes helped them during the test (i.e., prediction awareness). These prediction aware learners were always also gender aware and in most cases also pattern aware (n = 6). There were 3 prediction aware learners who were unaware of the pattern. As in Study 1, we observed that gender awareness was a prerequisite for developing other kinds of awareness.
The role of awareness, aptitudes, and instruction type in L2 processing

5.5.2 Eye movement results

For analyzing the eye-tracking data in Study 2, we used identical procedure as for Study 1.

Altogether 24.41% of all frames were missing or irrelevant (18.12% missing, 6.29% irrelevant), and these were not included in the analyses. Before analyzing the data, we excluded all trials in which more than 50% of the frames were missing or irrelevant. Using this procedure, 14.55% of the trials were excluded. There were no participants who lost more than 70% of the trials, so all participants were included in the analysis. We excluded exception trials because these items did not follow the dominant agreement pattern, and as a result, could have been processed differently.

First, we analyzed gaze data to see if learners overall showed predictive determiner processing. This was not the case, and in step 1 of the analysis, no initial significant time clusters were found.

Next, we looked at the effects of instruction type, i.e. whether rehearsing input influenced determiner processing. We found a significant effect of instruction type in a time cluster between 800 and 1200 ms after the onset of the determiner-adjective phase (Figure 5.5; Table 5.6). This effect implies that the two instruction groups showed different eye movement patterns. However, it is not entirely clear whether this difference was caused by the presence of predictive processing in the rehearsal group, or by the reversed patterns of divergence between the same- and different-gender trials in the silent group (Figure 5.5).

Figure 5.4. Overview of learners’ different levels of awareness as reported during debriefing (Study 2)
Figure 5.5. Proportion of learners’ looks towards the target image over time in the same-gender and different-gender trials, split by instruction type, and the time cluster during which the effect of instruction was observed (shaded area).

Table 5.6. Results of the cluster-based permutation analysis – effects of instruction type

<table>
<thead>
<tr>
<th>Time cluster</th>
<th>Sum statistic</th>
<th>Time range in ms</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.98</td>
<td>800 – 1200</td>
<td>0.015</td>
</tr>
</tbody>
</table>

We found that none of the different levels of awareness (gender awareness, pattern awareness, and prediction awareness) modulated learners’ predictive processing in this data sample. As Table 5.7 indicates, we found some clusters in step 1 of the analysis, but these did not reach significance in step 2. Unlike in Study 1, we found no evidence of processing based on prediction awareness. It is possible that this effect was not replicated due to the lack of power, given that there were only 9 learners who reported prediction awareness.
The role of awareness, aptitudes, and instruction type in L2 processing

Table 5.7. Results of the cluster-based permutation analysis – effects of learners’ awareness levels on predictive processing

<table>
<thead>
<tr>
<th>Time cluster</th>
<th>Sum statistic</th>
<th>Time range in ms</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender awareness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-7.25</td>
<td>800 – 950</td>
<td>0.16</td>
</tr>
<tr>
<td>Pattern awareness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.47</td>
<td>1450 – 1500</td>
<td>0.37</td>
</tr>
<tr>
<td>Prediction awareness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No clusters detected</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We found no links between determiner processing and different memory measures (LLAMA D and noun test). However, there was a small effect of analytical ability between 650 and 900 ms after the onset of the determiner-adjective phase, such that learners’ prediction score increased with lower analytical ability (Table 5.8).

Table 5.8. Results of the cluster-based permutation analysis – effects of aptitudes on learners’ predictive processing

<table>
<thead>
<tr>
<th>Time cluster</th>
<th>Sum statistic</th>
<th>Time range in ms</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLAMA D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7.82</td>
<td>0 – 150</td>
<td>0.14</td>
</tr>
<tr>
<td>2</td>
<td>2.29</td>
<td>1100 – 1150</td>
<td>0.38</td>
</tr>
<tr>
<td>Noun test score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No clusters detected</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IQ – Analytical ability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9.8</td>
<td>350 – 550</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>-17.07</td>
<td>650 – 900</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Finally, we checked whether any of the aptitude measures were predictive of different awareness levels, and we found no significant effects.

5.5.3 Reaction times (RTs)
We analyzed RTs in the same way as in Study 1. We found no overall speed advantage, i.e. there was no main effect of trial type \( (b = -49.71, SE = 46.48, t = -1.07, p = .28, 95\% CI [-142.67, 43.24]) \). Also, there was no evidence that the
instruction type that learners had received (rehearsal vs. silent) modulated their speed gains, i.e. the interaction between trial type and instruction type was not significant ($b = -26.39, SE = 51.05, t = -0.52, p = .61, 95\% CI [-128.49, 75.71])

As in Study 1, we found no evidence that gender awareness and pattern awareness contributed to learners’ speed advantage, i.e. there were no interactions between trial type and gender awareness ($b = 2.85, SE = 58.63, t = 0.05, p = .96, 95\% CI [-114.41, 120.1]) and between trial type and pattern awareness ($b = -9.55, SE = 56.68, t = -0.17, p = .87, 95\% CI [-122.92, 103.81])

When it comes to the role of prediction awareness, we observed a tendency that learners who were prediction aware gained some speed advantage on the different-gender trials (Table 5.9). However, this tendency, visible in the interaction between trial type and prediction awareness, was not significant, possibly due to a small number of prediction aware learners in this dataset ($b = -125.5, SE = 73.06, t = -1.72, p = .09, 95\% CI [-271.62, 20.61]).

Table 5.9. Overview of learners’ mean RTs and their standard deviations (SD), split by trial type and prediction awareness

<table>
<thead>
<tr>
<th>Prediction awareness</th>
<th>Different-gender trials</th>
<th>Same-gender trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aware ($n = 9$)</td>
<td>1860 (204)</td>
<td>2012 (92)</td>
</tr>
<tr>
<td>Unaware ($n = 41$)</td>
<td>2281 (224)</td>
<td>2308 (237)</td>
</tr>
</tbody>
</table>

When it comes to the role of different aptitudes, we found that none of the aptitude measures modulated learners’ speed advantage. The following interactions were not significant: trial type x LLAMA D ($b = -2.27, SE = 1.9, t = -1.19, p = .23, 95\% CI [-6.07, 1.54])$, trial type x IQ ($b = 8.1, SE = 16.05, t = 0.5, p = .61, 95\% CI [-40.19, 23.99])$, and trial type x noun test score ($b = -2.61, SE = 4.74, t = -0.55, p = .58, 95\% CI [-12.09, 6.87])

However, we found that the noun test performance ($b = 18.3, SE = 5.09, t = 3.6, p < .001, 95\% CI [8.13, 28.48])$ and the IQ scores ($b = -53.92, SE = 20.72, t = -2.6, p = .009, 95\% CI [-95.36, -12.48])$ predicted the speed with which learners could choose the correct image in general, but this effect did not differ in the same-gender and different-gender trials.
5.6 Discussion

The two studies presented in this chapter investigated the role of learners’ awareness, cognitive aptitudes, and instruction type in explaining differences in L2 learners’ ability to process determiners online. As noted in Roberts (2012), research so far has mostly focused on the role of L2 proficiency (e.g., Dussias et al., 2013; Dussias et al., 2014; Jackson, 2008; Sagarra & Herschensohn, 2010) and L1 characteristics (e.g., Dussias et al., 2013; Foucart & Frenck-Mestre, 2011; Sabourin & Stowe, 2008) in L2 processing, while other potentially important factors have received very little or no attention in research.

One of the main goals of this chapter was to investigate the role of learners’ awareness in the ability to develop predictive processing of determiners. The potential role that awareness may play in learners’ or native speakers’ online processing is very important to investigate given that most research implicitly assumes that using time-sensitive measures (e.g., ERPs, eye movements, RTs) may guarantee that the observed processing is implicit. However, as warned by several researchers, using a particular measure of processing cannot guarantee that the processing observed is explicit or implicit (Godfroid & Winke, 2015; Morgan-Short, Faretta-Stutenberg, & Bartlet-Hsu, 2015).

In our studies, we thoroughly debriefed participants, and we observed several levels of awareness among them: gender awareness, pattern awareness, and prediction awareness (i.e., learners’ awareness of their own predictive processing). The results demonstrated that only prediction aware learners showed signs of predictive processing of determiners, both in terms of eye movements and speed advantage, while gender awareness seems to have been a prerequisite for predictive awareness to develop.

Our data have important methodological implications for online processing research. They suggest that every study needs to establish whether the processing observed among L2 learners or native speakers is really implicit, or potentially results from some sort of task-awareness, i.e. from explicit and strategic processes. Furthermore, it is important to look at whether participants’ awareness of particular aspects of the target structure or task-related awareness can explain individual differences in processing. This is why thorough debriefing of participants is essential.

Online processing studies frequently report debriefing at the level of asking learners about the purpose of the experiment, which may be insufficient to detect more subtle levels of awareness. For instance, Dussias et al. (2013) report that “during debriefing, participants confirmed they were not aware that the focus of the experiment was grammatical gender” (p. 365). If researchers in this study asked participants only about the general purpose of the experiment and did not probe deeper, there may have been some levels of awareness among learners that went unobserved. For example, in the two studies...
presented in this chapter, none of the 150 learners were aware of the general focus of the experiment. Yet, 25 learners reported task-related prediction awareness in the form of a feeling that determiners were sometimes helpful in choosing the correct answer more quickly during the test. This suggests that thorough debriefing is needed to detect deeper levels of learners’ awareness, which in turn can help researchers to better understand the processing patterns observed among learners or native speakers.

The question that remains is whether the predictive processing that we observed in this study is completely strategic and intentional. One possibility would be that learners developed strategic prediction awareness early on in the eye-tracking test, and this helped them show predictive processing of determiners. The other possibility is that learners developed implicit intuitions based on which they used determiners to anticipate the correct noun, and that these intuitions gradually led them to become aware of their own predictive processing. Based on the results of our studies, it is not possible to draw clear conclusions about this, but we have some reasons to argue that it is unlikely that determiner-based prediction in this study was fully strategic. First, learners who reported prediction awareness during debriefing described it as a sort of feeling they had during the test, rather than a strategy that they intentionally adopted from the start. This observation, although not systematically investigated in our studies, may tentatively indicate that prediction awareness developed as a result of learners becoming aware of their own predictive processing. Second, because of the nature of the target structure used in our study, we argue that knowledge deeper than prediction awareness was needed to show predictive processing. A simple comparison with the study of Andringa and Curcic (in preparation) can be used to illustrate this point. Andringa and Curcic (in preparation) provided some evidence of prediction by unaware participants, but later provision of a brief explicit instruction radically improved learners’ predictive processing, which is indicative of strategic processing. In that study, it was possible for strategic awareness to have such a strong, quick effect on predictive processing because the target structure used was very simple: once learners were told that preposition announces an animate object in a sentence, it was easy for them to show predictive processing without any deeper knowledge of the structure. However, such strategic performance would not be possible with determiner-based prediction, where even if learners were told that they can predict based on the determiner, they would not be able to do it without having well entrenched knowledge of all different determiner-noun combinations. Therefore, even if one would assume that all processing found in our study was deliberate and strategic from the very start, learners needed substantial knowledge of the target structure in order to show prediction, and prediction awareness in itself would not have been enough.
Coming back to our question of whether awareness of structures plays a role in processing, the results of this chapter indicate that learners' gender awareness and pattern awareness could not directly translate into enhanced predictive processing. However, we did find gender awareness to be a prerequisite for learners to develop prediction awareness; there was only 1 learner (out of 25) who was prediction aware without also being aware of the gender distinction in the language. This may imply that awareness of certain aspects of L2 structures may indirectly help processing by enabling learners to develop other levels of awareness that are more directly associated with enhanced online processing. Our results are in line with previous research that suggested that awareness of L2 structures or explicit L2 instruction may not directly affect online processing (e.g., Andringa & Curcic, 2015). The reason may be that mere awareness of L2 structures is not enough, and more consolidated knowledge is needed before learners can come to process L2 structures in nativelike ways. This might also be the reason why learners with higher L2 proficiency are more likely to process L2 structures in nativelike ways (e.g., Dussias et al., 2013; Jackson, 2008; Sagarra & Herschensohn, 2010), and why abundant naturalistic exposure is more beneficial for online processing than explicit classroom instruction (e.g., Frenck-Mestre, 2002; Pliatsikas & Marinis, 2013).

Our next goal was to investigate the role of cognitive aptitudes in online processing: working memory and analytical ability. While several studies have identified working memory as a cognitive aptitude that may contribute to learners' online processing (e.g., Dussias & Piñar, 2010; Havik et al., 2009; Williams, 2006), we found no direct evidence of a relation between working memory and predictive processing of determiners. This means that there was no evidence that learners with better working memory could process determiners better. However, in Study 1 we observed a tendency that learners' working memory capacity predicted whether they would develop prediction awareness, and as discussed earlier, this level of awareness was associated with predictive processing. This might imply that working memory was indirectly involved in online processing via awareness. With respect to the role of analytical ability, in Study 2 we found a very small effect, but in the opposite direction. Weaker analytical ability led to better online processing, which is difficult to explain. However, given that this effect was not strong and was observed in a relatively brief time cluster of 250 ms, it should be treated with caution.

Finally, we looked at whether language instruction with rehearsal may lead to better online processing of determiners. Our hypothesis was based on several findings in the literature: that rehearsal helps to keep input longer in working memory (Baddeley & Hitch, 1974), and that working memory sometimes predicts online processing (e.g., Dussias & Piñar, 2010; Havik et al,
2009; Williams, 2006). Also, the study presented in Chapter 4 has found that rehearsal fosters learning of determiner-noun combinations, so we hypothesized that it may also lead to better determiner-based prediction. In the present study, we observed significant differences in the online processing of learners who rehearsed during instruction and those who did not. The effects observed mainly came from the fact that between 800 and 1200 ms after the onset of the determiner, learners in the rehearsal group showed a stronger preference for the target image in the different-gender trials than in the same-gender trials, whereas for the silent group, the opposite was true: they showed a stronger preference for the target image in the same-gender trials than in the different-gender trials. While this result points to some instruction-related effect on processing, it does not provide evidence that instruction with rehearsal led to improved processing. Also, we did not observe that learners in the rehearsal group gained determiner-based speed advantage when responding to different gender test items. This study is need of replication, preferably with a larger sample size in order to be able to draw clearer conclusions about the effects of rehearsal on processing of determiners.

There are a few limitations of the study that need to be mentioned. First, we looked at the online processing of beginner learners, who are generally less likely to show processing due to insufficient proficiency levels. While it is very interesting and relevant that we found evidence of predictive processing of determiners even after a very brief exposure to language input, the limited length of exposure very likely influenced the amount of online processing we observed, and possibly limited the statistical power of the study, especially when we looked at how different levels of awareness affected processing. Also, we based our variable of awareness on learners’ retrospective verbal reports. This method might have led to an underestimation of the number of learners who actually developed some awareness of the target structure or of their own predictions because they may not have been able to verbalize it (Rebuschat et al., 2015).

This study contributes to the body of research that investigates factors that may explain differences in learners’ online processing. We investigated the relation between L2 processing on the one hand, and awareness, aptitudes, and instruction types on the other hand because these factors have not received much attention in research so far. More research in this area will be needed that focuses on different aspects of these factors (e.g., different aptitudes, different kinds of awareness, different instruction types), and relates them to L2 processing of different language structures.
Chapter 6: General discussion and conclusions

In this dissertation, we investigated two factors involved in explaining differences in adult language learning: input characteristics and learners' cognitive aptitudes. We were interested in how input reliability affects learning the determiner-noun agreement pattern, as well as whether individual differences in learning could be explained by learners' working memory, analytical ability, and statistical learning ability. We also looked at aptitude by treatment interactions, i.e. whether involvement of the aptitude measures may differ for learning from reliable and unreliable input.

In the studies that we presented in this dissertation, learners were briefly exposed to a miniature language based on Fijian and the target determiner-noun agreement pattern. We employed several measures of language knowledge: oral grammaticality judgment tasks (GJTs), a production task, and online processing tasks with eye-tracking. The learners in our studies differed substantially in whether they developed awareness of the target pattern, so we always took awareness into account when interpreting results pertaining to our research questions.

In this chapter, we will summarize and discuss the main findings of this dissertation through several subsections. In this way we will answer the research questions introduced in Chapter 1, and we will give some suggestions for further research. After this, we will discuss some general methodological observations and insights that came from our studies. Finally, in the last subsection, we will bring together potential practical implications of this dissertation.

6.1 How reliability of the input affected language learning

Given that input and its characteristics play a crucial role in language learning, it is surprising that relatively few studies have investigated how subtle changes and manipulations of input properties may or may not influence learning of particular L2 structures. This kind of research can be invaluable both theoretically – to understand how input features affect the process of learning, and practically – to improve language instruction and facilitate learning. This dissertation contributes to the small body of research in the L2 field that used carefully controlled experimental designs and introduced subtle changes in the input to see how input characteristics may affect language learning (e.g., research on the effectiveness of skewed vs. balanced input: Casenhiser &
In Chapters 2 and 3, we presented two studies that investigated how input reliability affects learning the determiner-noun agreement pattern in adult learners. Findings confirmed the important role of input characteristics in L2 learning (Bybee, 2008), and showed that subtle variations in input reliability may lead to differences in pattern learning, which has been suggested by several researchers (e.g., DeKeyser, 1994; Hulstijn & De Graaff, 1994). Both studies provided clear evidence that unreliable input with exceptions disturbed learning of determiner-noun agreement for learners who had already developed awareness of this pattern. We observed that aware learners who had heard exceptions in the input were less likely to rely on the pattern they had consciously noticed during grammaticality judgment tasks, and were also considerably less accurate in their judgments. However, it is important to note that these findings are based on group comparisons. In both studies, there were substantial differences between learners: some aware learners were strongly disturbed by exceptions in the input, while some learners were immune to the negative effects of unreliable input, and retained very high accuracies on the grammaticality judgment tasks. In child L1 acquisition of grammatical structures, such as plural marking and clitic placement, it has also been found that unreliable input can disturb learning (e.g., Costa et al., 2015; Miller & Schmitt, 2012). Although we observed similar effects of unreliable input for adult L2 learning, it should be stressed that L1 learning is mostly considered to be implicit, while in our studies, disturbance of learning was observed among aware learners.

We found no evidence that unreliable input disturbed learning among participants who were unaware of the target pattern. In Chapter 2, there was no evidence that input had any effect on the test performance of pattern unaware learners. However, in Chapter 3, where learners received somewhat longer exposure to the target pattern, we found positive effects of unreliable input: learners exposed to unreliable input memorized determiner-noun combinations from the input – both regular items and exceptions – better than learners who had heard fully reliable input. These positive effects of unreliable input are especially interesting given that these learners were unaware of the unreliability of the input, i.e. the existence of any pattern or exceptions to the pattern. However, as this dissertation shows, longer exposure to language input may be needed for these effects to become visible. The fact that we found positive effects of unreliable input in unaware learners, and negative effects of unreliable input in aware learners shows that the effects that input characteristics have on language learning may interact with other factors, such as awareness.
Studies in this dissertation made some of the first steps in investigating how input reliability affects L2 pattern learning. Further research on this topic is needed that looks at different structures and uses even longer instructional treatments, preferably with delayed posttests. Also, in order to gain more insight into how and when input reliability affects learning, future studies should look at how differential proportions of regular input vs. exceptions may or may not affect positive or negative effects on learning, as well as when exceptions should be introduced in learning in order to achieve optimal learning outcomes. Finally, the possibility that input reliability effects may depend on other factors, such as learners’ awareness or instruction type, should be taken into consideration and further investigated in future research.

6.2 Aptitudes and pattern learning

Most researchers agree on the complex, multidimensional nature of aptitude (DeKeyser & Koeth, 2011; Dornyei, 2006; Skehan, 2002; Robinson, 2002, 2005a). This implies 1) that we should talk about multiple aptitudes, rather than a single aptitude (DeKeyser & Koeth, 2011; Dornyei, 2006), and 2) that involvement of particular aptitudes in language learning may interact in complex ways with other factors, such as instruction, age, motivation, learning stages, etc. (DeKeyser 2012; Robinson, 2002, 2005a; Skehan, 2002). Throughout this dissertation, we investigated the relation between learning of the determiner-noun agreement pattern and several cognitive aptitudes: working memory, analytical ability, and statistical learning ability. We also tried to do justice to the complex nature of aptitudes by looking at aptitude by treatment interactions.

We employed several tests to measure aptitudes, and the tests we used sometimes differed across the studies presented in this dissertation. In all studies, we used the LLAMA D test (Meara, 2005) as a measure of learners’ verbal working memory. Also, learners in all studies did a noun test, which could be regarded as a measure of rote memory. In some studies, we included additional working memory measures: the non-verbal digit span task (Olsthoorn et al., 2014) (in the long exposure study) and the non-word repetition task (NWR) (in the rehearsal study; De Jong & Van der Leij, 1999). While working memory measures (i.e., LLAMA D, digit span task, NWR task) required learners to retain verbal or non-verbal stimuli for very short periods of time, and later reproduce them or distinguish them from novel stimuli, the noun test tapped into rote memory; it measured learners’ ability to make longer-term connections between words and their referents. To measure learners’ analytical ability, we used the non-verbal matrix reasoning subtest of the WAIS IQ battery in all studies (Wechsler, 2008). Additionally, in the short exposure study, we used the LLAMA F test (Meara, 2005), tapping into learners’
verbal analytical ability. This study also included the serial reaction time task (SRT; Kaufman et al., 2010), as a measure of learners’ non-verbal statistical learning ability. The reason why our aptitude measures somewhat differed across studies was that we sometimes dropped the measures that did not predict learning (e.g., LLAMA F and SRT), and we sometimes added new measures that would potentially better match with the nature of the learning we investigated (e.g., NWR in the study on rehearsal).

Before we proceed with a discussion of the links between aptitude measures and learning observed in this dissertation, it needs to be noted that learners’ performance on the GJT with novel nouns and their awareness reports were indicators of pattern knowledge. Learners’ performance in the GJT with familiar items could be based either on pattern knowledge or memory, or a mixture of these two. This means that we did not necessarily expect to observe the same involvement of aptitude measures in learners’ performance on the two GJTs.

Across the different studies presented in Chapters 2, 3, and 4, fairly consistent links emerged between learners’ performance on the GJT with familiar nouns and particular aptitudes. We found that learners who performed better on LLAMA D and noun test were more likely to also achieve higher accuracies on the GJT with familiar nouns; this was the case in all three studies (Table 6.1). Additionally, in the long exposure study, links with the digit span task have also been found. This implies that learners’ memory capacities generally predicted their performance on the GJT with familiar nouns. Also, in the study on rehearsal, learners’ production of regular items was predicted by their noun test scores, whereas learners’ memory-based production of exceptions was predicted by the NWR measure of working memory. These links with memory are not surprising given that both GJT with familiar nouns and the production test at least partially tapped into learners’ memorizing of the input, i.e. item-based knowledge. We also found some links between learners’ analytical abilities (i.e., IQ test) and their performance on the GJT with familiar nouns. These links were present in the short exposure and rehearsal studies, but not in the long exposure study. The involvement of aptitude measures in the GJT with familiar nouns (i.e., both memory and analytical ability) suggests that learners’ performance on this task was likely based on both memory and pattern learning.

When it comes to learners’ performance on the GJT with novel nouns, we observed a more complex, less consistent picture, and also several aptitude by treatment (group) interactions. Generally speaking, there were links between pattern learning (as measured by the GJT with novel nouns) and both analytical abilities and memory measures, but the exact involvement of the different measures and their interactions with group varied substantially between different studies (Table 6.1). The fact that participants often found the GJT with
novel nouns difficult and confusing may have partially contributed to the inconsistent findings we observed. Across the three studies, we found some links between memory measures and learners’ accuracies on the GJT with novel nouns, and these links were sometimes modulated by reliability of input. For instance, in the short exposure study, higher scores on LLAMA D predicted higher accuracies on the novel GJT for reliable input learners. Similarly, in the long exposure study, digit span task predicted reliable input learners’ performance. This means that memory measures were involved in pattern learning, but that their involvement was often stronger when learners had received fully reliable input. This may imply that under unreliable input conditions, links between pattern learning and memory become less visible, possibly due to the fact that unreliable input disturbs pattern learning for many learners. When it comes to the role of analytical ability in learners’ performance on the GJT with novel nouns, we observed a similar interaction with input reliability in the long exposure study: high IQ scores were related to better pattern learning from reliable but not unreliable input. However, this result should be taken with caution given that it did not appear consistently across the different studies. Finally, we found that learners’ awareness of the pattern was predicted by LLAMA D scores (in the short and long exposure studies), noun test scores (in the rehearsal study), and also IQ scores (in the short exposure study).

**Table 6.1. Overview of the involvement of memory measures (M) and analytical ability measures (AA) in familiar and novel GJT performance and pattern awareness for the short exposure, long exposure, and rehearsal studies.**

<table>
<thead>
<tr>
<th></th>
<th>GJT familiar nouns</th>
<th>GJT novel nouns</th>
<th>Pattern awareness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>AA</td>
<td>M</td>
</tr>
<tr>
<td><strong>Short exposure study</strong></td>
<td>LLAMA D &lt;br&gt;Noun test</td>
<td>IQ</td>
<td>LLAMA D x group</td>
</tr>
<tr>
<td><strong>Long exposure study</strong></td>
<td>LLAMA D</td>
<td>none</td>
<td>LLAMA D x group</td>
</tr>
<tr>
<td><strong>Rehearsal study</strong></td>
<td>LLAMA D &lt;br&gt;Noun test</td>
<td>IQ</td>
<td>Noun test</td>
</tr>
</tbody>
</table>
Generally speaking, the results presented above imply that both memory and analytical ability are involved in pattern learning, which is in line with previous research (for links between working memory and pattern learning, see Denhovska et al., 2015; Robinson, 2005b; Sagarra & Herschensohn, 2010; for links between analytical ability and pattern learning, see Erlam, 2005; Harley & Hart, 1997; Robinson, 2005b). Our results also point to the existence of complex aptitude by treatment interactions, which is in line with the complex, multidimensional views of aptitude (e.g., DeKeyser, 2012). This may point to the need for research that does justice to the complex nature of aptitude.

Given that in Chapters 2 and 3 we found relatively consistent links between working memory and pattern learning (in line with Denhovska et al., 2015; Kempe et al., 2010; Robinson, 2005b; Sagarra & Herschensohn, 2010, etc.), in Chapter 4 we investigated whether instruction with input rehearsal could foster pattern learning. Considering that rehearsal helps to keep input longer in short-term memory (Baddeley & Hitch, 1974), we expected that it may improve pattern learning via memory (found in Ellis & Sinclair, 1996). This hypothesis was largely confirmed. There was some evidence that rehearsal helped learners who were unaware of the pattern to memorize determiner-noun combinations better. Also, when learners were aware of the pattern, rehearsal helped them to apply it better during the GJT with familiar nouns. There was also a tendency that rehearsal was beneficial for aware learners’ performance on the GJT with novel nouns. These findings may be seen as a confirmation of the involvement of working memory and beneficial effects of rehearsal in L2 pattern learning (in line with Ellis & Sinclair, 1996). They also suggest that knowledge of which aptitudes are involved in language learning may have important practical implications for adapting instruction in such ways that learning is improved.

Finally, several words of caution are in place when it comes to the aptitude-related results of this dissertation and aptitude research in general. Although in this dissertation we found some consistent links between aptitudes and learning, we also observed many inconsistencies and unclear results. First, the observed inconsistencies may have been caused by slight differences in the input (i.e., short exposure vs. long exposure vs. exposure with rehearsal). As we have already discussed, even within the studies, we observed that involvement of aptitudes depended on the kind of input learners had heard.

Second, the inconsistencies in aptitude involvement found in this dissertation are probably not surprising if we look at the field of aptitude research in general, where it is quite common to find inconsistencies in links between aptitudes and learning, as well as inconsistencies in the definition of aptitudes and choice of aptitude measures. This brings us to the more general problems in aptitude-related research. Even after many years of extensive research, it is still difficult to draw solid conclusions about how language
learning and aptitudes are linked, and there are still many uncertainties and open questions in the field. One of the reasons is certainly the complex, multidimensional nature of aptitude, and complex interactions between aptitudes and other factors. Several researchers have therefore suggested that it is not sufficient to look at simple correlations between aptitudes and learning outcomes, and that in research designs, we should do more justice to the complex interactions between aptitudes and other factors, such as learning stages, contextual, motivational, and other variables (e.g., DeKeyser, 2012; Dörnyei, 2009; Robinson, 2002; Skehan, 2002).

However, some challenges associated with such research need to be pointed out. The main one may lie in the difficulty to make clear and informed hypotheses about the processes and aptitudes that may underlie learning of particular structures, in particular contexts, at particular stages of learning. Such knowledge is probably an essential first step in designing experimental research that properly investigates aptitudes and does justice to their complexity. With this in mind, it could be said that aptitude researchers often end up in a vicious circle, where they lack knowledge and experimental evidence to form clear hypotheses and conduct high quality aptitude research, and at the same time, they need to conduct research in order to be able to develop knowledge and form hypotheses.

While we tried to at least partially go beyond simple correlations between aptitudes and learning, and investigate aptitude by treatment interactions in this dissertation, we found it hard to formulate clear, theoretically motivated hypotheses, and to make principled decisions about which aptitude tests to use. A consistent and clear picture simply did not emerge from the previous research, and this is probably exactly why this dissertation has not provided fully consistent results when it comes to the involvement of aptitude in learning.

### 6.3 Individual differences in L2 processing

The studies presented in Chapter 5 of this dissertation investigated the involvement of aptitudes, different levels of awareness, and instruction types in explaining differences in learners’ online processing of determiners, as measured by the visual world eye-tracking paradigm. While the role of L2 proficiency and L1 characteristics in learners’ L2 processing has been confirmed through experimental research (e.g., Dussias et al., 2013; Foucart & French-Mestre, 2011; Sabourin & Stowe, 2008), other factors that may influence L2 processing have been hardly investigated (Roberts, 2012).

Most studies on online L2 processing looked at whether learners at certain proficiency levels can process their L2 structures in nativelike ways, but very few studies employed designs with a brief language exposure or instruction to
see if these may quickly lead to enhanced processing (e.g., Andringa & Curcic, 2015; Batterink & Neville, 2013; Davidson & Indefrey, 2009; Marsden et al., 2013; Morgan-Short et al., 2010). For instance, Andringa and Curcic (2015) used the visual world paradigm with eye-tracking and found no signs of predictive processing of the target differential object marking structure after a brief explicit or implicit exposure to a miniature language. However, several other studies used ERP measures, and indicated that more nativelike processing of grammatical structures in L2 can develop after very brief instructional treatments (e.g., Batterink & Neville, 2013; Davidson & Indefrey, 2009; Morgan-Short et al., 2010). These studies investigated how different instructional treatments (e.g., explicit vs. implicit instruction) affected development of online processing, but did not look at some other potentially relevant factors, such as learners’ awareness developed during exposure or cognitive aptitudes.

In this dissertation, we used the visual world paradigm with eye-tracking (similar to Andringa & Curcic, 2015), and we found evidence that online processing of L2 patterns can develop even after a very brief exposure to a novel language. This is in line with several studies that have also found evidence of online processing of L2 patterns after a brief exposure (e.g., Batterink & Neville, 2013; Davidson & Indefrey, 2009; Morgan-Short et al., 2010). However, it needs to be noted that these studies employed ERP measures, which unlike visual world paradigm include ungrammatical stimuli and may be more likely to trigger learners’ awareness of the target pattern or task-related awareness.

In this dissertation, we found three levels of awareness among learners: gender awareness (i.e., awareness of the gender distinction), pattern awareness (i.e., awareness of the agreement pattern between nouns and determiners), and prediction awareness (i.e., learners’ awareness of using determiners during the online processing test). We were interested in whether each of these levels of awareness could explain individual differences in learners’ predictive processing of determiners. The results clearly suggest that gender awareness was a prerequisite for the development of predictive processing of determiners, and that only prediction aware learners were able to process determiners online and gain a speed advantage based on that processing. However, based on our data, it is not completely clear whether prediction awareness fostered predictive processing, or learners developed this level of awareness because they noticed and contemplated about their own prediction ability. In any case, there are indications that the processing we observed was not a purely strategic processing based on a trick. Learners who reported prediction awareness mostly reported it as a kind of feeling that they developed during the test, rather than a strategy. Additionally, using determiners as a predictive cue required solid knowledge of different determiner-noun combinations, and a
mere awareness that determiners could be used to choose the correct image would not have been enough.

We also investigated involvement of cognitive aptitudes in learners' ability to develop predictive processing of determiners. Research on this topic is still scarce (Roberts, 2012), and the existing research on the role of working memory in L2 processing has produced mixed results (Havik et al., 2009), with some studies finding links between working memory and L2 processing (e.g., Dussias & Piñar, 2010; Havik, et al., 2009; Williams, 2006) and others finding no such links (e.g., Felser & Roberts, 2007; Juffs, 2004, 2005; Traxler, 2007). In this dissertation, we found no clear evidence that working memory and analytical ability directly fostered learners' online processing of determiners. However, given that working memory tended to predict the development of learners' prediction awareness, and only prediction aware learners used determiners as a cue in their processing, we may assume that working memory indirectly fostered predictive processing of determiners.

Finally, we found that the instruction that learners received (i.e., with or without rehearsal) affected their online processing. Learners who rehearsed input during exposure showed different processing patterns from the ones who did not rehearse. However, based on our results it is not straightforward whether this difference can really be interpreted as improved, more nativelike processing. Further research on this topic, preferably with large samples, is needed to see if clearer results would emerge.

6.4 Awareness-related differences

Researchers in the SLA field have been well aware that different instructional treatments or measures of learners' L2 knowledge may trigger awareness of L2 structures (Andringa & Rebuschat, 2015; Norris & Ortega, 2000). For instance, as pointed out in Andringa and Rebuschat (2015), grammaticality judgment tasks and even online processing measures such as ERPs may trigger learners' awareness of the target structures due to the use of ungrammatical items. This dissertation showed that learners differed substantially in whether or not they developed awareness during the experiment, and these differences have also been observed within subgroups of learners who received identical language exposure. Learners differed in what exactly they became aware of (e.g., gender distinction, target pattern, using determiners to predict the correct answer during the eye-tracking test, etc.), as well as when they developed the awareness. Some learners developed awareness already during the exposure, while for others awareness was triggered during one of the two GJTs, or during the debriefing. The finding that some learners were even aware of their own online processing of determiners during the eye-tracking task is quite interesting given that the visual world eye-tracking paradigm is considered a
very unobtrusive measure, not likely to trigger awareness because it does not feature ungrammatical items (Andringa & Rebuschat, 2015). Our findings are in line with the views that emphasize the absence of a one-to-one relationship between measures of language processing and explicit or implicit processing (Godfroid & Winke, 2015; Morgan-Short et al., 2015), which also may imply that using a certain measure does not guarantee that learners will or will not develop awareness during the task.

The results of this dissertation point to awareness as a very important source of individual differences in language learning. They also suggest that looking at awareness-related variables can help language researchers understand differences in learning outcomes better, and can also help understand how effects of certain variables may differ for aware vs. unaware learners. For instance, by looking at the effects of input reliability in both aware and unaware learners, we observed that exceptions in the input disturbed learning of aware learners but fostered learning by unaware learners. Without taking into account learners’ awareness, this finding would not have surfaced. Similarly, in the first online processing study presented in Chapter 5, we found that learners as a group used determiners as a cue to predict the coming noun. However, once we took awareness differences into consideration, it became clear that this effect was entirely driven by learners who were aware of their own predictive processing during the test. We found no evidence that learners without this awareness showed predictive processing. Finally, by taking awareness differences into account in this dissertation, we were able to speculate on how awareness and implicit learning may develop over time. We observed that with longer exposure, implicit learning was no longer visible, possibly because it was overridden by memory effects.

6.5 Methodological considerations

In this section we will consider several valuable methodological insights that resulted from this dissertation, relating to the use of artificial languages, debriefing, the use of GJT's, and considering exposure length and moment of testing in research. For each of these points, we will discuss how they may be addressed in the future research.

In this dissertation, we used a miniature semi-artificial language to investigate language learning in adults. This paradigm offers a number of advantages over naturalistic research and the use of existing languages. Its major advantage is that it allows for controlling the amount and type of input that learners are exposed to. This is especially important when investigating individual differences in language learning because in this line of research, the goal is to isolate effects of particular variables (e.g., aptitudes, motivation, etc.), while at the same time controlling for as many other variables that may also
contribute to differences between learners. For instance, if the goal is to assess how learners’ working memory ability influences their learning outcomes, one would ideally want all learners to have received the same amount and kind of language input because such a design can help to isolate the effects of working memory. While very strong links between working memory and learning may be observed even with many interfering variables present, more subtle links may be obscured by interfering variables. Using a non-existing or a very rare language also eliminates the need of carefully controlling for learners’ knowledge of the languages that may resemble the target language used in research.

However, several things need to be taken into account when using miniature artificial languages. Whenever possible, using artificial languages that feature learners’ L1 or known L2 vocabularies should be avoided because this can make learners aware that they are learning a non-existing language, or may even involve unlearning, rather than learning a language. Paradigms in which learners are exposed to artificial languages that feature their L1 or L2 vocabulary may direct learners’ attention to the structures under investigation, thereby triggering metalinguistic reflection and awareness. The artificial language used in this dissertation featured novel vocabulary that learners were not familiar with, and thanks to this approach, participants could not know that the language they learned was not an existing language. They often assumed it was a Slavic language given that the materials were recorded by a native speaker of Serbian. Another essential first step when using miniature artificial languages is to ensure that learning effects are observed (Rogers, Revesz, & Rebuschat, 2015). Ensuring that a language is learnable on the basis of the exposure provided requires carefully balancing the complexity of its structures, complexity of the input, and the length of the exposure. The language needs to be carefully tested through pilot studies to make sure that learning effects will be observed. This is something that Rogers et al. (2015) noted in their research, and we also noted it in this dissertation. Namely, the initial version of the language materials in this dissertation included transitive sentences as well. However, a pilot study showed that this exposure was too complex for the target pattern to be learned by the participants in the time we gave them, probably because it featured additional verbs that took learners’ attention away from the target structure. This observation required us to simplify the language in order to make it learnable within a limited amount of exposure.

Careful debriefing of participants in this dissertation proved very informative. Learners were able to express very well what awareness they had developed, when this awareness occurred, as well as how they approached different tests, and what they relied on during the tests. These data helped us understand and interpret our data better. Given that awareness can affect learners’ behavior during experimental tasks so dramatically, no study using
adult participants can afford not to conduct careful and thorough debriefing (for a good discussion of the importance of debriefing, see Rebuschat et al., 2015). Such in-depth debriefing does require time and careful coding, but it is certainly an invaluable tool for better understanding the results, and interpreting the differences observed between learners.

Next, we should address some insights related to the use of GJT in investigating pattern learning in SLA. As noted in Rebuschat et al. (2015), and as confirmed in this dissertation, GJT may trigger learners’ awareness of the target patterns. We found that some learners indeed developed awareness of the agreement pattern during one of the two GJT. This means that GJT is certainly not ideal tasks for investigating implicit knowledge. We observed that the use of GJT with novel items that learners had never been exposed to is especially inappropriate. Although these GJT are supposed to tap into learners’ ability to generalize a pattern to novel items, in this dissertation they did not work well for detecting implicit learning. During this task, many learners did not feel comfortable enough to guess or use their intuition, and they instead either came up with a wrong rule and relied on that, or they focused on nouns only, rather than nouns in combination with determiners. In our analyses, we chose not to exclude learners who relied on a wrong rule or had a wrong focus because it was not clear what triggered these wrong sources of reliance. One possibility is that learners who relied on a wrong rule or had a wrong focus did develop some knowledge of the target structure, but confusion led them down a wrong path. The other possibility is that these sources of reliance simply resulted from learners’ lack of knowledge. GJT with familiar items were much easier for learners given that they needed to judge the appropriateness of the items that they had already heard before. Even learners unaware of the pattern felt comfortable enough to rely on guessing, intuition, or simply memory.

This dissertation presents a potentially fruitful paradigm for investigating early stages of implicit learning. Namely, in Chapter 2, we found that learners who were not aware of the pattern overgeneralized the dominant pattern to exceptions, instead of judging items based on their memory for the input, which can be interpreted as evidence of pattern knowledge. This paradigm may prove to be a very fruitful one given that it relies on learners’ judging of familiar items that they had heard in the exposure, rather than requiring them to judge novel items, which learners often find very uncomfortable and confusing.

This dissertation also shows the value of manipulating input length when using exposure paradigms. Such manipulations may enable researchers to map learning trajectories and indirectly investigate how learning progresses in time. Also, the moment at which learners’ knowledge of target structures is tested needs to be carefully considered because tests administered at different moments in time may tap into different stages of learning, where different processes may be at play. For instance, in our studies, we observed that testing
after short exposure gave evidence of implicit pattern learning, whereas testing after longer exposure showed effects of memorization. This may imply that with prolonged exposure to input, memory effects start, and signs of implicit pattern knowledge get overridden by memory.

Finally, we would like to point to the importance of replications in research in order to better understand and validate previously obtained research results (for detailed discussions, see Mackey, 2012; Porte, 2010; Polio & Gass, 1997). For instance, in this dissertation, replication gave us more certainty about the effects of unreliable input on pattern learning in aware learners, as well as about the involvement of aptitude measures in learning.

6.6 Practical implications

Based on the results of this dissertation, it is difficult to see immediate implications for teaching. However, there are certainly a number of intriguing findings and methodological observations that deserve to be addressed in future experimental research and research in naturalistic settings.

When it comes to the implications for language instruction, the findings of this dissertation suggest that with explicit types of instruction, where learners are expected to become aware of patterns, it may be beneficial to initially reduce the amount of exceptions in the input, and introduce them later in the learning process, rather than straight from the beginning. However, before these implications are firmly drawn, more research is needed to see if smaller amount of exceptions in the input or introducing exceptions later in the learning process diminishes the negative effects that unreliable input has on explicit learning. Furthermore, results of this dissertation point to some beneficial effects of rehearsal in L2 learning, where learners who rehearse language input may be able to better memorize regular items and exceptions from the input. Also, for aware learners, rehearsal seems to help better application of the pattern and better memorizing of exceptions. However, although rehearsal is a fairly intensive and focused activity, it is possible that less intensive, but meaningful production of target structures may lead to similar positive effects. Further research is needed to find out if this is the case, and if rehearsal activities also have an added value when combined with context-based types of instruction.

Finally, this dissertation offers important methodological insights for future research. Our results stress the importance of using thorough debriefing of participants and considering potential differences in their awareness in order to understand results better. Also, we show that even when using relatively unobtrusive measures of language processing, researchers need to be careful when drawing conclusions about the nature of the observed processing. In other words, using unobtrusive measures does not guarantee that learners’
performance is based on implicit processes. Again, thorough debriefing should be used to help in drawing conclusions on what kind of processing has been measured.
## Appendices

### Appendix A: Target language words used in the exposure

<table>
<thead>
<tr>
<th>Word</th>
<th>Meaning</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nouns – consistent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. ganeuk</td>
<td>boy</td>
<td>masc.</td>
</tr>
<tr>
<td>2. taganuk</td>
<td>man</td>
<td>masc.</td>
</tr>
<tr>
<td>3. karetuk</td>
<td>bird</td>
<td>masc.</td>
</tr>
<tr>
<td>4. dawauk</td>
<td>cat</td>
<td>masc.</td>
</tr>
<tr>
<td>5. isleuk</td>
<td>horse</td>
<td>masc.</td>
</tr>
<tr>
<td>6. oseuk</td>
<td>ball</td>
<td>masc.</td>
</tr>
<tr>
<td>7. vualuk</td>
<td>car</td>
<td>masc.</td>
</tr>
<tr>
<td>8. kumiuk</td>
<td>window</td>
<td>masc.</td>
</tr>
<tr>
<td>9. maramis</td>
<td>girl</td>
<td>fem.</td>
</tr>
<tr>
<td>10. jalewis</td>
<td>woman</td>
<td>fem.</td>
</tr>
<tr>
<td>11. burogis</td>
<td>dog</td>
<td>fem.</td>
</tr>
<tr>
<td>12. araris</td>
<td>mouse</td>
<td>fem.</td>
</tr>
<tr>
<td>13. sitois</td>
<td>rabbit</td>
<td>fem.</td>
</tr>
<tr>
<td>14. eulis</td>
<td>bicycle</td>
<td>fem.</td>
</tr>
<tr>
<td>15. salis</td>
<td>chair</td>
<td>fem.</td>
</tr>
<tr>
<td>16. ikais</td>
<td>computer</td>
<td>fem.</td>
</tr>
</tbody>
</table>

| **Nouns – inconsistent**                    |         |          |
| 1. touk       | goat    | masc. (reliable input) fem. (unreliable input) |
| 2. senikauk   | book    | masc. (reliable input) fem. (unreliable input) |
| 3. vonuis     | turtle  | fem. (reliable input) masc. (unreliable input) |
| 4. ligis      | box     | masc. (unreliable input) |

| Verbs (3rd person singular)               |         |          |
| 1. sisilit | swims | -        |
| 2. gadet   | runs  | -        |
| 3. bulat   | sleeps | -        |
| 4. na      | is    | -        |
### Adjectives

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>duka</td>
<td>broken/torn</td>
<td>-</td>
</tr>
<tr>
<td>2.</td>
<td>kamusu</td>
<td>dirty</td>
<td>-</td>
</tr>
<tr>
<td>3.</td>
<td>vinake</td>
<td>yellow</td>
<td>-</td>
</tr>
<tr>
<td>4.</td>
<td>matene</td>
<td>blue</td>
<td>-</td>
</tr>
<tr>
<td>5.</td>
<td>galile</td>
<td>red</td>
<td>-</td>
</tr>
<tr>
<td>6.</td>
<td>ramase</td>
<td>green</td>
<td>-</td>
</tr>
</tbody>
</table>

### Determiners

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>lep</td>
<td>the</td>
</tr>
<tr>
<td>2.</td>
<td>ris</td>
<td>the</td>
</tr>
</tbody>
</table>
Appendix B: Target structure exposure items

1) Determiner + noun items

*Items with consistent nouns*
Appendices

Items with inconsistent nouns

Lep/ris touk
“The goat”

Lep/ris senikauf
“The book”

Ris/lep vonuis
“The turtle”

Ris/lep ligis
“The box”

Items with consistent nouns

Lep ramase ganeuk
“The green boy”

Lep vinake karetuk
“The yellow bird”

Lep galile vuakuk
“The red car”

Lep matene oseuk
“The blue ball”

Ris galile maramis
“The red girl”

Ris vinake araris
“The yellow mouse”

Ris matene eulis
“The blue bicycle”

Ris ramase ikais
“The green computer”

2) Determiner + adjective_{color} + noun items

These are example items. All 20 nouns were combined with each of the four color adjectives.
Items with inconsistent nouns

Lep/ris vinake touk
"The yellow goat"

Lep/ris galile senikauk
"The red book"

Ris/lep ramase vonuis
"The green turtle"

Ris/lep matene ligis
"The blue box"

3) Intransitive sentences: determiner + animate noun + verb\textsubscript{intransitive}

Items with consistent nouns

Lep ganeuk gadet.
"The boy is running."

Lep ganeuk sisilit.
"The boy is swimming."

Lep taganuk gadet.
"The man is running."

Lep taganuk bulat.
"The man is sleeping."

Lep dawauk gadet.
"The cat is running."

Lep dawauk bulat.
"The cat is sleeping."

Lep karetuk gadet.
"The bird is running."

Lep karetuk bulat.
"The bird is sleeping."

Lep iseleuk gadet.
"The horse is running."

Lep iseleuk sisilit.
"The horse is swimming."

Ris maramis gadet.
"The girl is running."

Ris maramis bulat.
"The girl is sleeping."
“The woman is running.”
“The woman is swimming.”
“The dog is running.”
“The dog is swimming.”
“The mouse is running.”
“The mouse is sleeping.”
“The rabbit is running.”
“The rabbit is sleeping.”
“The goat is running.”
“The goat is sleeping.”
“The turtle is swimming.”
“The turtle is sleeping.”
4) Intransitive sentences: determiner + inanimate noun + is + adjective

*Items with consistent nouns*

- **Lep oseuk na duka.** “The ball is torn.”
- **Lep oseuk na kamusu.** “The ball is dirty.”
- **Lep vuakuk na duka.** “The car is broken.”
- **Lep vuakuk na kamusu.** “The car is dirty.”
- **Lep kumiuk na duka.** “The window is broken.”
- **Lep kumiuk na kamusu.** “The window is dirty.”
- **Ris eulis na duka.** “The bicycle is broken.”
- **Ris eulis na kamusu.** “The bicycle is dirty.”
- **Ris salis na duka.** “The chair is broken.”
- **Ris salis na kamusu.** “The chair is dirty.”
- **Ris ikais na duka.** “The computer is broken.”
- **Ris ikais na kamusu.** “The computer is dirty.”
Appendices

*Items with inconsistent nouns*

Lep/ris seniakaun *na duka.*
“The book is torn.”

Lep/ris seniakaun *na kamusu.*
“The book is dirty.”

Ris/lep ligis *na duka.*
“The box is torn.”

Ris/lep ligis *na kamusu.*
“The box is dirty.”
Appendix C: GJT with novel nouns

<table>
<thead>
<tr>
<th>Noun type</th>
<th>-uk nouns</th>
<th>-is nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatical</td>
<td>1. Lep drauk</td>
<td>1. Ris namuis</td>
</tr>
<tr>
<td></td>
<td>“The lion”</td>
<td>“The pig”</td>
</tr>
<tr>
<td></td>
<td>2. Lep batauk</td>
<td>2. Ris tinis</td>
</tr>
<tr>
<td></td>
<td>“The frog”</td>
<td>“The fish”</td>
</tr>
<tr>
<td></td>
<td>3. Lep oteluk</td>
<td>3. Ris gonis</td>
</tr>
<tr>
<td></td>
<td>“The monkey”</td>
<td>“The elephant”</td>
</tr>
<tr>
<td></td>
<td>4. Lep gusuk</td>
<td>4. Ris manis</td>
</tr>
<tr>
<td></td>
<td>“The cake”</td>
<td>“The ice-cream”</td>
</tr>
<tr>
<td></td>
<td>5. Lep ivoluk</td>
<td>5. Ris katenis</td>
</tr>
<tr>
<td></td>
<td>“The table”</td>
<td>“The sandwich”</td>
</tr>
<tr>
<td></td>
<td>6. Lep vosuk</td>
<td>6. Ris polois</td>
</tr>
<tr>
<td></td>
<td>“The bag”</td>
<td>“The glasses”</td>
</tr>
<tr>
<td></td>
<td>7. Lep basuk</td>
<td>7. Ris kolis</td>
</tr>
<tr>
<td></td>
<td>“The pen”</td>
<td>“The fork”</td>
</tr>
<tr>
<td></td>
<td>8. Lep urouk</td>
<td>8. Ris yamis</td>
</tr>
<tr>
<td></td>
<td>“The candle”</td>
<td>“The umbrella”</td>
</tr>
<tr>
<td></td>
<td>9. Lep soruk</td>
<td>9. Ris vatuis</td>
</tr>
<tr>
<td></td>
<td>“The apple”</td>
<td>“The glass”</td>
</tr>
<tr>
<td></td>
<td>10. Lep teveluk</td>
<td>10. Ris daligis</td>
</tr>
<tr>
<td></td>
<td>“The bottle”</td>
<td>“The guitar”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grammaticality</th>
<th>Ungrammatical</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. *Ris drauk</td>
<td>1. *Lep namuis</td>
</tr>
<tr>
<td>“The lion”</td>
<td>“The pig”</td>
</tr>
<tr>
<td>2. *Ris batauk</td>
<td>2. *Lep tinis</td>
</tr>
<tr>
<td>“The frog”</td>
<td>“The fish”</td>
</tr>
<tr>
<td>3. *Ris oteluk</td>
<td>3. *Ris gonis</td>
</tr>
<tr>
<td>“The monkey”</td>
<td>“The elephant”</td>
</tr>
<tr>
<td>“The cake”</td>
<td>“The ice-cream”</td>
</tr>
<tr>
<td>5. *Ris ivoluk</td>
<td>5. *Lep katenis</td>
</tr>
<tr>
<td>“The table”</td>
<td>“The sandwich”</td>
</tr>
<tr>
<td>“The bag”</td>
<td>“The glasses”</td>
</tr>
<tr>
<td>7. *Ris basuk</td>
<td>7. *Lep kolis</td>
</tr>
<tr>
<td>“The pen”</td>
<td>“The fork”</td>
</tr>
<tr>
<td>8. *Ris urouk</td>
<td>8. *Lep yamis</td>
</tr>
<tr>
<td>“The candle”</td>
<td>“The umbrella”</td>
</tr>
<tr>
<td>“The apple”</td>
<td>“The glass”</td>
</tr>
<tr>
<td>“The bottle”</td>
<td>“The guitar”</td>
</tr>
</tbody>
</table>
### Appendix D: GJT with familiar nouns

<table>
<thead>
<tr>
<th>Noun type</th>
<th>-uk nouns</th>
<th>-is nouns</th>
</tr>
</thead>
</table>

Appendix E: Test with eye-tracking

<table>
<thead>
<tr>
<th>Trial Type</th>
<th>Target gender</th>
<th>Masculine</th>
<th>Feminine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Same gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. Lep ramase karetuk/kareleuk</td>
<td>&quot;The yellow bird/horse&quot;</td>
<td>&quot;The green mouse/rabbit&quot;</td>
</tr>
<tr>
<td></td>
<td>2. Lep matene dawauk/kareleuk</td>
<td>&quot;The blue cat/bird&quot;</td>
<td>&quot;The blue dog/mouse&quot;</td>
</tr>
<tr>
<td></td>
<td>3. Lep galiite kaleleuk/dawauk</td>
<td>&quot;The red horse/cat&quot;</td>
<td>&quot;The red rabbit/dog&quot;</td>
</tr>
<tr>
<td></td>
<td>4. Lep ramase oseuk/vuakuk</td>
<td>&quot;The green ball/car&quot;</td>
<td>&quot;The green bicycle/chair&quot;</td>
</tr>
<tr>
<td></td>
<td>5. Lep matene vuakuk/kareleuk</td>
<td>&quot;The blue car/window&quot;</td>
<td>&quot;The blue chair/computer&quot;</td>
</tr>
<tr>
<td></td>
<td>6. Lep galiite kumuek/oseuk</td>
<td>&quot;The red window/ball&quot;</td>
<td>&quot;The red computer/bicycle&quot;</td>
</tr>
<tr>
<td></td>
<td>7. Lep vinake taganuk/ganeuk</td>
<td>&quot;The yellow man/boy&quot;</td>
<td>&quot;The yellow girl/woman&quot;</td>
</tr>
<tr>
<td></td>
<td>8. Lep vinake ganeuk/taganuk</td>
<td>&quot;The yellow boy/man&quot;</td>
<td>&quot;The yellow woman/girl&quot;</td>
</tr>
<tr>
<td></td>
<td>Inconsistent items:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9. Lep/Ris vinake touk/kareleuk</td>
<td>&quot;The yellow goat/bird&quot;</td>
<td>&quot;The yellow girl/woman&quot;</td>
</tr>
<tr>
<td></td>
<td>10. Lep/Ris ramase senikeuk/kareleuk</td>
<td>&quot;The green book/window&quot;</td>
<td>&quot;The red turtle/mouse&quot;</td>
</tr>
<tr>
<td></td>
<td>Different gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. Lep galiite kareleuk/sitois</td>
<td>&quot;The red bird/rabbit&quot;</td>
<td>&quot;The red mouse/bird&quot;</td>
</tr>
<tr>
<td></td>
<td>2. Lep ramase dawauk/burogis</td>
<td>&quot;The green cat/dog&quot;</td>
<td>&quot;The green dog/cat&quot;</td>
</tr>
<tr>
<td></td>
<td>3. Lep matene iseluek/araris</td>
<td>&quot;The blue horse/mouse&quot;</td>
<td>&quot;The blue rabbit/horse&quot;</td>
</tr>
<tr>
<td></td>
<td>4. Lep galiite oseuk/ikais</td>
<td>&quot;The red ball/computer&quot;</td>
<td>&quot;The red bicycle/ball&quot;</td>
</tr>
<tr>
<td></td>
<td>5. Lep ramase vuakuk/eulis</td>
<td>&quot;The yellow car/bicycle&quot;</td>
<td>&quot;The green chair/car&quot;</td>
</tr>
<tr>
<td></td>
<td>6. Lep matene kumuek/salais</td>
<td>&quot;The blue window/chair&quot;</td>
<td>&quot;The blue computer/window&quot;</td>
</tr>
<tr>
<td></td>
<td>7. Lep vinake taganuk/maramis</td>
<td>&quot;The yellow man/girl&quot;</td>
<td>&quot;The yellow girl/boy&quot;</td>
</tr>
<tr>
<td></td>
<td>8. Lep vinake ganeuk/jalewis</td>
<td>&quot;The yellow boy/woman&quot;</td>
<td>&quot;The yellow woman/man&quot;</td>
</tr>
<tr>
<td></td>
<td>Inconsistent items:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9. Lep/Ris ramase touk/sitois</td>
<td>&quot;The green goat/rabbit&quot;</td>
<td>&quot;The red box/ball&quot;</td>
</tr>
<tr>
<td></td>
<td>10. Lep/Ris vinake senikeuk/salais</td>
<td>&quot;The yellow book/chair&quot;</td>
<td>&quot;The blue turtle/horse&quot;</td>
</tr>
</tbody>
</table>
Appendices
References


References


References


References


Wong, W., & VanPatten, B. (2003). The evidence is IN: Drills are OUT. *Foreign Language Annals, 36,* 403–423.
References
Summary
Explaining differences in adult second language learning: The role of language input characteristics and learners’ cognitive aptitudes

Adults differ substantially in how successful they are in learning a new language. One of the questions that have intrigued second language researchers is how those differences can be explained. The value of investigating and explaining individual differences between language learners is of both theoretical and practical nature. Research on individual differences helps researchers to understand what processes underlie language learning and how it progresses. And even more importantly, this knowledge has the potential to help in improving language instruction.

Broadly speaking, this dissertation investigated two sources that may lead to differences in language learning success: the role of language input that learners hear during their language acquisition and their cognitive aptitudes for language learning.

Throughout the dissertation, we investigated Dutch adult learners’ acquisition of a determiner-noun agreement structure in a semi-artificial language based on Fijian, which is an Austronesian language spoken in Fiji. The target determiner-noun agreement pattern was the following: feminine determiner *ris* preceded nouns ending in *–is* (e.g., *ris burogis* — “the dog”) and masculine determiner *lep* preceded nouns ending in *–uk* (e.g., *lep iseleuk* — “the horse”). Learners in each study received brief auditory exposure to the novel language and the target structure. The exposure started with a noun learning phase, followed by a noun test, and once the nouns were learned, participants were exposed to the target structure. After this, learners did a test with eye-tracking to see if they had learned to use determiners to predict the upcoming nouns (data presented in Chapter 5), a production test that measured their productive knowledge of the agreement pattern (only in Chapter 4 study), and two grammaticality judgment tasks that measured learners’ receptive knowledge of the target pattern (data presented in Chapters 2, 3, and 4). After these tests, we thoroughly debriefed participants in order to find out if they had become aware of the determiner-noun agreement pattern, which we later used to better understand and interpret the results throughout the dissertation.

In Chapter 2, we investigated how unreliable input with exceptions affected learners’ acquisition of determiner-noun agreement. We were also interested in whether several cognitive aptitudes (i.e., working memory, analytical ability,
and statistical learning ability) could explain individual differences in learning outcomes between learners.

In usage-based approaches to second language acquisition (SLA), it has been observed that different characteristics of language input (e.g., salience, frequency, reliability, etc.) may influence learning outcomes (Bybee, 2008). Some of these observations have been investigated in carefully controlled experimental studies, where researchers modified language input in order to assess how input characteristics affected learning outcomes. How input reliability (i.e., whether or not a rule contains exceptions) affects language learning has been investigated only in first (L1), but not in second language (L2) learning. The L1 literature has suggested that variable or unreliable input can delay child acquisition of grammatical structures (e.g., Costa et al., 2015; Miller, 2007; Miller & Schmitt, 2012), but the L2 field lacks systematic research in this area. In order to find out how input reliability affected L2 pattern learning, we divided learners in our study in two groups. For half of the learners in the study, the determiner-noun agreement pattern was fully regular (i.e., reliable input group), and for half of them it contained exceptions (i.e., unreliable input group).

Accumulated research on the relation between cognitive aptitudes and linguistic pattern learning has provided some evidence that working memory, analytical ability, and statistical learning ability are among the aptitudes that are involved in pattern learning. However, there has also been an increasing recognition that aptitudes interact with a number of other variables, such as instruction, motivation, age, etc., which may require looking beyond simple relations between aptitude and learning, and looking at complex interactions between aptitudes and other variables (DeKeyser, 2012). This is why in Chapter 2 we also tried to look at whether the involvement of working memory, analytical ability, and statistical learning ability may differ when learners are exposed to reliable vs. unreliable input.

Chapter 2 results confirmed our expectation that unreliable input would negatively affect pattern learning. It also provided evidence of implicit pattern learning, and complex interactions between aptitudes and instruction.

Based on learners’ performance on the grammaticality judgment tasks (GJTs), the study provided evidence that unreliable input disturbed learning, but only for those learners who developed awareness of the target structure. Apparently, unreliable input made many aware learners feel less confident about the pattern they had noticed, which in turn led to lower accuracies on the GJTs. However, not all aware learners were equally affected by exceptions in the input, and some learners who heard exceptions during their exposure were equally accurate as learners who had heard fully reliable input. For learners who remained unaware of the target pattern, we found no evidence that input reliability negatively affected their learning. Learning rates of these learners
were much lower. However, we did find some evidence of implicit learning of the target pattern in that the unaware learners overgeneralized the dominant pattern to exceptions.

Learning outcomes in the study were predicted by learners’ working memory, rote memory ability (i.e., the measure obtained from the noun test), and analytical ability. Also, an interaction between working memory and input reliability was found, such that working memory predicted pattern learning, but only from fully reliable input. Both working memory and analytical ability predicted if learners would develop awareness of the target pattern.

Chapter 3 reports on a study in which we used the same design as in the study described above, but this time learners were given 60% longer exposure to the target agreement pattern. This enabled us to see if we could replicate the effects of unreliable input and the involvement of aptitudes in learning that we had found in Chapter 2. An important motivation was also to investigate if longer exposure would lead to higher rates of awareness of the target pattern among learners, to higher levels of implicit pattern learning, or both.

Interestingly, we replicated some of the findings from Chapter 2, whereas for others we observed a rather different pattern of results. For the aware learners, the effects of unreliable input on pattern learning were fully replicated, which means that exceptions in the input disturbed performance of aware learners, and similar individual differences in this effect were found. However, for the unaware learners, we no longer found evidence of implicit pattern learning. We only found evidence of memorizing determiner-noun combinations among unaware learners, and unreliable input seemed to foster better memorization for these learners. We found that longer exposure led to disappearance of implicit knowledge evidence, and that learners who were exposed to the language for longer time were also more likely to develop awareness of the target pattern. This result at first sight seems in line with Cleeremans (2011) and Bialystok (1989, 1994), who suggested that accumulated implicit knowledge or increased opportunities for analyzing input may gradually lead learners to become aware of the pattern. However, in this study we found that the differences in awareness in the short and long exposure studies stemmed from differences in working memory in the two groups of learners, which means that we cannot provide clear evidence of Cleeremans’s and Bialystok’s assumptions. However, our findings do imply two things: 1) that effects of input reliability may interact with exposure length, and 2) what kind of knowledge is assessed (e.g., implicit pattern knowledge or memorization of input) may depend on language exposure and the moment of testing, which is an important methodological consideration for future studies that try to assess language learning after brief exposure.

Links between aptitudes and learning outcomes were replicated, but not fully. Generally speaking, in both Chapter 2 and Chapter 3, we found some links
between learning on the one hand and learners’ working memory, rote memory, and analytical ability on the other hand. We also found evidence that working memory interacted with input reliability, i.e. it predicted pattern learning from reliable but not unreliable input. In Chapter 3, we also observed a similar interaction between analytical ability and input reliability. Both studies showed that working memory predicted whether learners would become aware of the target pattern. Analytical ability was also predictive of awareness in Chapter 2, but not Chapter 3.

In Chapter 4, we report on another study where we investigated if exposure with rehearsal may promote learning of the determiner-noun agreement pattern. Our hypothesis that this may be the case was based on several insights from previous research. First, previous research found links between pattern learning and working memory, and in Chapters 2 and 3 we also observed consistent links between learning of the target structure and working memory. Second, according to one of the influential models of working memory, rehearsal is a component of working memory that helps to keep input longer in working memory, which in turn may help to promote longer-term learning (Baddeley & Hitch, 1974).

Chapter 4 study found some evidence of the beneficial effects of rehearsal for learning the determiner-noun agreement structure. When learners were aware of the target pattern and rehearsed language input during their exposure, they could better acquire determiner-noun combinations and memorize exceptions. This implies that rehearsal promotes better memorization, and also potentially better application of the pattern by aware learners. When learners were unaware of the pattern, there was some evidence that rehearsal helped them to better memorize determiner-noun combinations. Interestingly and paradoxically, while rehearsal may have helped aware learners to apply the pattern better, exposure without rehearsal was more conducive to developing awareness of the pattern, possibly because the activity of rehearsing may take cognitive resources away from analyzing the input. In this study, we again found links between learning outcomes on the one hand and working memory, rote memory, and analytical ability on the other hand.

In Chapter 5 of this dissertation, we looked at online processing of determiners in learners who participated in the three previously reported studies. We wanted to know whether learners could develop the ability to use determiners to rapidly predict the coming noun, as measured by the visual world paradigm with eye-tracking. Such processing has been widely attested among native speakers of languages with determiners (e.g., Brouwer et al., 2010; Dussias et al., 2013) and also among L2 learners, depending on their L2 proficiency and the characteristics of their L1 (e.g., Dussias et al., 2013). This dissertation investigated if several other factors that were rarely investigated in relation to online L2 processing could predict learners’ ability to develop
predictive processing of determiners: awareness, cognitive aptitudes (i.e., working memory and analytical ability), and instruction type (with or without rehearsal).

Through extensive debriefing, we found three levels of awareness among our learners: gender awareness, agreement pattern awareness, and prediction awareness (i.e., awareness of the usefulness of determiners during the test). Only prediction aware learners showed predictive processing of determiners and gained a speed advantage thanks to this processing. While gender awareness was not related to prediction as such, it was found to be a prerequisite for developing prediction awareness. One may be tempted to think that predictive processing of determiners found in this dissertation was the result of learners’ purely strategic behavior, but we argue that it is very unlikely. First, learners reported prediction awareness as a feeling that they developed during the test rather than a strategy that they adopted on purpose. Second, the predictive processing of determiners required extensive knowledge of determiner-noun combinations, and a mere awareness that determiners could be used during the test would not have been sufficient. Therefore, a more likely explanation seems to be that learners developed prediction awareness in the course of the test, as a result of noticing their own predicative processing.

We found no evidence that either working memory or analytical ability directly fostered learners’ ability to show predictive processing of determiners. However, the data did suggest that working memory may be indirectly involved, given that it tended to predict whether or not learners would develop prediction awareness.

As for the effects of instruction type, we found no straightforward evidence that rehearsal led to improved determiner-based predictive processing among learners, although we did observe that the rehearsal and silent groups differed significantly in their processing patterns.

This dissertation offers several methodological insights that may be invaluable for future experimental SLA research. First, our results stress the importance and value of conducting thorough post-experimental debriefing of learners in order to better interpret research results, and check what kind of knowledge or processing the previously administered tests really measured. Second, we offer a potentially promising paradigm for assessing implicit knowledge using items that learners had heard during exposure. This paradigm may allow researchers to detect subtle signs of implicit pattern learning that may not be measurable using grammaticality judgment tasks with novel items given that novel items may create too much confusion among learners to enable detecting implicit knowledge. Last but not least, given that the length of exposure and the timing of testing were found to heavily influence what learning processes were observed, this dissertation points to the need to
carefully consider them when designing experimental research with the goal of finding a particular type of learning.

Broadly speaking, this dissertation provides some novel insights on how characteristics of language input and learners' cognitive aptitudes may be involved in the learning of novel linguistic patterns. We show that reliability of language input that learners hear during their learning process can dramatically affect learning outcomes in either positive or negative ways, depending on how learning proceeds. Our research also provides some further support for the involvement of working memory and analytical ability in L2 pattern learning, and also for the existence of complex aptitude by treatment interactions, where involvement of aptitudes in learning may depend on the input that learners are exposed to. Finally, the outcomes of this dissertation provide some important insights with respect to how learners' awareness at different levels may affect language learning and outcomes, both in terms of learners' performance on offline, untimed tasks and their rapid online processing of language.
Samenvatting

Tweedetaalverwerving bij volwassenen: De rol van taalaanbod kenmerken en de cognitieve capaciteiten van taalleerders

Volwassenen verschillen aanzienlijk in hoe succesvol ze zijn in het leren van een nieuwe taal. Een van de vragen die de onderzoekers intrigeert is hoe die verschillen verklaard kunnen worden. Het onderzoeken en verklaren van individuele verschillen in taalleersucces is zowel theoretisch als praktisch relevant. Onderzoek naar individuele verschillen helpt ons begrijpen welke processen ten grondslag liggen aan het taalleerproces en hoe dat proces verloopt. Belangrijker is dat deze kennis ook kan helpen het taalonderwijs te verbeteren.

In dit proefschrift zijn twee bronnen onderzocht die tot verschillen kunnen leiden in het taalleersucces: de eigenschappen van het taalaanbod dat leerders tijdens het taalverwervingsproces te horen krijgen en hun cognitieve capaciteiten voor het leren van een tweede taal.

Centraal in dit proefschrift staat de verwerving van lidwoord-substantiecongruentie in een semi-kunstmatige taal gebaseerd op het Fiji, een Austronesische taal die op de Fiji eilanden wordt gesproken. De deelnemers waren volwassen moedertaalsprekers van het Nederlands. Het congruentiepatroon tussen het lidwoord en het substantief was als volgt: het vrouwelijk lidwoord ris ging vooraf aan zelfstandige naamwoorden die op -is eindigen (bijv. ris burogis - "de hond") terwijl het mannelijke lidwoord lep voorafging aan zelfstandige naamwoorden die op -uk eindigen (bijv. lep iseleuk - "het paard"). In elke studie werden deelnemers auditief aan de nieuwe taal en de doelstructuur blootgesteld. De blootstelling was kort en begon met het leren van de zelfstandige naamwoorden, gevolgd door een woordenschattoets. Zodra de deelnemers de zelfstandige naamwoorden hadden geleerd, werden ze aan de artificiële taal met de doelstructuur blootgesteld. Hierna deden ze een toets met eye-tracking om te zien of ze lidwoorden hadden leren gebruiken om de zelfstandige naamwoorden te voorspellen (resultaten gepresenteerd in Hoofdstuk 5), een productietoets die hun productieve kennis van het congruentiepatroon mat (gepresenteerd in Hoofdstuk 4), en twee grammaticaliteitsoordeeltaken die de receptieve kennis van het doelpatroon maten (resultaten gepresenteerd in Hoofdstuk 2, 3 en 4). Na deze toetsen zijn de deelnemers uitvoerig geïnterviewd om na te gaan of ze zich bewust waren van de congruentie tussen lidwoord en zelfstandig naamwoord. Deze
informatie is gebruikt om de resultaten van het proefschrift beter te begrijpen en te interpreteren.

In Hoofdstuk 2 is onderzocht hoe onbetrouwbare taalaanbod (door uitzonderingen op de regel) de verwerving van het congruentiepatroon beïnvloedde. Verder is nagegaan of bepaalde cognitieve capaciteiten (d.w.z. werkgeheugen, analytisch vermogen en statistisch leervermogen) individuele verschillen in taalleersucces konden verklaren.

Tweedetaalverwervingsonderzoek heeft aangetoond dat bepaalde kenmerken van het taalaanbod (bijvoorbeeld opvallendheid, frequentie en betrouwbaarheid van de doelstructuur) de verwerving kunnen beïnvloeden (Bybee, 2008). In sommige gevallen is gebruik gemaakt van zorgvuldig gecontroleerde experimentele studies, waarbij onderzoekers het taalaanbod hebben gemanipuleerd om te beoordelen hoe bepaalde aanbodkenmerken de leerresultaten beïnvloeden. De effecten van betrouwbaarheid (d.w.z. of er al dan niet uitzonderingen op een regel zijn) zijn wel onderzocht voor eerstetaalverwerving (T1), maar niet voor tweedetaalverwerving (T2). De T1-literatuur suggereert dat variabel of onbetrouwbaar taalaanbod de verwerving van grammaticale structuren bij kinderen kan vertragen (bijv. Costa et al., 2015; Miller, 2007; Miller & Schmitt, 2012), maar het T2-veld ontbeert systematisch onderzoek op dit gebied. Om erachter te komen of de betrouwbaarheid van taalaanbod van invloed was op het leren van een nieuw T2-patroon, zijn de deelnemers in ons onderzoek in twee groepen verdeeld. Voor de helft van de deelnemers was het congruentiepatroon volkomen regelmatig (de betrouwbaar aanbodgroep), voor de andere helft waren er uitzonderingen in het patroon (de onbetrouwbaar aanbodgroep).

Onderzoek naar de relatie tussen cognitieve capaciteiten en het leren van taalpatronen heeft enig bewijs opgeleverd dat een goed werkgeheugen, analytisch vermogen en statistisch leervermogen het leren van taalstructuren bevorderen. Het wordt echter steeds duidelijker dat er sprake is van interactie tussen cognitieve capaciteiten en een aantal andere variabelen, zoals instructie, motivatie en leeftijd. Dat wil zeggen dat we misschien verder moeten kijken dan alleen naar eenvoudige relaties tussen cognitieve capaciteiten en taalleersucces, en dat we meer aandacht aan deze interacties moeten besteden (DeKeyser, 2012). Om die reden hebben we in Hoofdstuk 2 ook onderzocht of de relatie tussen werkgeheugen, analytisch vermogen en statistisch leervermogen aan de ene kant en het leren van een nieuw taalpatroon aan de andere kant afhing van het type taalaanbod (betrouwbaar of onbetrouwbaar) waaraan de deelnemers werden blootgesteld.

De resultaten van Hoofdstuk 2 bevestigden onze verwachting dat een onbetrouwbare taalaanbod een negatief effect heeft op de verwerving van het congruentiepatroon. De studie leverde ook aanwijzingen op dat de verwerving
Samenvatting

van dit patroon onbewust plaatsvond, en er was bewijs voor interacties tussen cognitieve capaciteiten en het taalaanbod.

De prestaties van de leerders op de grammaticaliteitsoordeeltaken lieten zien dat het onbetrouwbare taalaanbod het leren hinderde, maar dit gold alleen voor die leerders die zich bewust werden van de doelstructuur. Blijkbaar zorgde het onbetrouwbare aanbod ervoor dat bewuste leerders minder vertrouwen hadden in het patroon dat ze hadden opgemerkt, wat tot een lagere accuratesse op de oordeeltaken leidde. Niet alle bewuste leerders werden echter evenzeer getroffen door de uitzonderingen in het aanbod. Sommige leerders die uitzonderingen hoorden in het taalaanbod waren even accuraat als de leerders die volledig betrouwbaar aanbod hadden gehoord. Voor leerders die het doelpatroon niet opmerkten, vonden we geen bewijs dat betrouwbaarheid het leren van de taal negatief beïnvloedde. De scores van deze leerders waren veel lager, maar er was wel enig bewijs van impliciet leren van de doelstructuur doordat de onbewuste leerders het dominante patroon overgeneraliseerden naar de uitzonderingen.

In deze studie werd taalleersucces voorspeld door de werkgeheugencapaciteit van de leerders, het vermogen om woordjes uit het hoofd te leren (op basis van de maat verkregen uit de woordenschattoets) en hun analytisch vermogen. Ook werd er een interactie vastgesteld tussen het werkgeheugen en de betrouwbaarheid van het taalaanbod: het werkgeheugen voorspelde het leren van het doelpatroon, maar dit was alleen het geval voor het leren van betrouwbaar taalaanbod. Ten slotte bleek dat zowel de werkgeheugencapaciteit als het analytisch vermogen voorspelden of leerders zich bewust werden van het doelpatroon.

Hoofdstuk 3 rapporteert over een studie waarin dezelfde onderzoeksopzet is gebruikt als in de hierboven beschreven studie, maar in deze studie werden de leerders zestig procent langer blootgesteld aan de doelstructuur. Dit stelde ons in staat om de resultaten van Hoofdstuk 2 te repliceren. Een belangrijk doel was ook om te onderzoeken of langere blootstelling aan de doelstructuur zou leiden tot een groter bewustzijn van het doelpatroon, tot hogere niveaus van impliciet leren, of tot beide.

Interessant is dat een aantal van de bevindingen uit Hoofdstuk 2 zijn gerepliceerd, terwijl we voor andere bevindingen duidelijk andere resultaten kregen. Voor de bewuste leerders werden de effecten van onbetrouwbaar aanbod op het leren van de doelstructuur volledig gerepliceerd. De uitzonderingen in het taalaanbod verstoorden de prestaties van de bewuste leerders. Ook vonden we weer dat niet alle bewuste leerders evenzeer getroffen werden door de uitzonderingen in het aanbod. Voor de onbewuste leerders vonden we echter geen bewijzen meer voor impliciet leren. In plaats daarvan vonden we dat langere blootstelling leidde tot betere opslag van de lidwoord-substantiefcombinaties en dat het onbetrouwbare taalaanbod het
Samenvatting

leren zelfs leek te bevorderen. Ook werden de deelnemers die langer aan het taalaanbod waren blootgesteld zich vaker bewust van het doelpatroon. Op het eerste gezicht lijkt dit resultaat in overeenstemming te zijn met Cleeremans (2011) en Bialystok (1989, 1994), die suggereerden dat geaccumuleerde impliciete kennis of toegenomen mogelijkheden voor het analyseren van input ertoe kunnen leiden dat leerders zich geleidelijk bewust worden van het patroon. Echter, in deze studie konden we de verschillen in bewustwording niet zonder meer toeschrijven aan de blootstellingsduur. Wel ontdekten we dat de verschillen in bewustzijn tussen de leerders met korte en lange blootstelling voortvloeiden uit verschillen in werkgeheugencapaciteit tussen de twee groepen leerders. Deze replicatiestudie leverde twee belangrijke inzichten op: 1) het effect van de betrouwbaarheid van het taalaanbod bleek afhankelijk van de blootstellingsduur, en 2) de lenglengte van de blootstelling leek gevolgen te hebben voor de aard van de gemeten kennis (korte blootstelling leidde tot impliciete patroonkennis en vergeneralisatie; lange blootstelling tot gememoriseerde kennis). Deze inzichten zijn van belang bij het ontwerpen van toekomstige studies die het leren van de taal na een korte blootstelling proberen te beoordelen.

De verbanden tussen cognitieve capaciteiten en taalleersucces werden gerepliceerd, maar niet volledig. Over het algemeen hebben we zowel in Hoofdstuk 2 als in Hoofdstuk 3 enkele verbanden gevonden tussen taalexers enerzijds en werkgeheugencapaciteit, het vermogen om woordjes uit het hoofd te leren en het analytisch vermogen van de leerders anderzijds. We hebben ook vergelijkbare patronen gevonden voor de interactie tussen werkgeheugen en taalaanbodbetrouwbaarheid: het werkgeheugen voorspelle wel patroonleren uit betrouwbaar, maar niet uit onbetrouwbaar taalaanbod. Verder lieten beide studies zien dat werkgeheugencapaciteit voorspelle of leerders zich bewust zouden worden van het doelpatroon. Analytisch vermogen was ook voorspellend voor het bewustzijn in Hoofdstuk 2, maar niet in Hoofdstuk 3.

In Hoofdstuk 4 rapporteren we over een andere studie waarin we onderzochten of blootstelling met herhaling het leren van lidwoordsubstantiecongruentie kan bevorderen. Onze hypothese dat dit het geval zou kunnen zijn was gebaseerd op verschillende inzichten uit eerder onderzoek. Ten eerste heeft eerder onderzoek verbanden gevonden tussen patroonleren en werkgeheugen; in Hoofdstuk 2 en 3 vonden we eveneens consistente verbanden tussen het leren van de doelstructuur en het werkgeheugen. Ten tweede is herhaling, volgens een invloedrijk werkgeheugenmodel, een aspect van het werkgeheugen dat helpt om het aanbod langer in het werkgeheugen te houden, wat op zijn beurt het van leren op de langere termijn bevordert (Baddeley & Hitch, 1974).

De studie in Hoofdstuk 4 levert enig bewijs voor de gunstige effecten van herhaling voor het leren van de congruentiestructuur. In deze studie werden
twee condities vergeleken: enkel blootstelling versus blootstelling plus herhaling. Bovendien werd niet alleen de passieve kennis gemeten, er werd ook een productietak toegevoegd. Deelnemers in de repetitieconditie die zich bewust werden van het doelpatroon waren beter in staat de lidwoord-substantiecombinaties te onthouden dan deelnemers in de conditie zonder herhaling en dat gold ook voor de uitzonderingen op de regel. Dit houdt in dat repetitie bevorderlijk is voor het memoriseren en mogelijk ook voor een betere toepassing van het patroon door bewuste leerders. Ook wanneer deelnemers zich niet bewust werden van het patroon was er enig bewijs dat de herhaling hun hielp om de lidwoord-substantiecombinaties beter te onthouden. Hoewel de herhaling de leerders mogelijk heeft geholpen om het patroon beter te kunnen toepassen, bleek paradoxaal genoeg dat blootstelling zonder herhaling juist bevorderlijk bleek voor het ontwikkelen van bewustzijn van het patroon (leerders in de conditie zonder herhaling werden zich vaker bewust van de doelstructuur), mogelijk omdat de activiteit van het herhalen analyse van het taalaanbod in de weg staat. In deze studie vonden we opnieuw verbanden tussen taaleren enerzijds en werkgeheugen, het vermogen om woordjes uit het hoofd te leren en analytisch vermogen anderzijds.

In hoofdstuk 5 hebben we gekeken naar de online verwerking van lidwoorden bij de leerders die aan de drie eerder gerapporteerde studies hebben deelgenomen. Het doel was om erachter te komen of de leerders het vermogen konden ontwikkelen om op basis van het lidwoord het erop volgende zelfstandig naamwoord te voorspellen. Dit is gedaan door de oogbeweging van de leerders te volgen in het zogenaamde 'visual world paradigm'. Dergelijke voorspellingen op basis van lidwoorden zijn vaak geobserveerd bij moedertaalsprekers van talen met lidwoorden (bijv. Brouwer et al., 2010; Dussias et al., 2013). Hetzelfde geldt voor tweedetaalleerders, maar dan is het afhankelijk van hun taalvaardigheid in de tweede taal en de kenmerken van hun eerste taal (bijv. Dussias et al., 2013). In dit proefschrift is onderzoek gedaan naar een aantal andere factoren met betrekking tot online tweedetaalverwerking die zelden zijn onderzocht: de rol van bewustzijn, cognitieve capaciteiten (d.w.z. werkgeheugen en analytisch vermogen) en het type instructie (met of zonder herhaling). Het doel was om na te gaan of er sprake is van een relatie tussen die factoren en het vermogen om lidwoorden voorspellend te leren gebruiken.

Met behulp van uitgebreide debriefing interviews werden drie niveaus van bewustzijn bij deelnemers vastgesteld: bewustzijn voor grammaticaal geslacht, bewustzijn voor de congruentie tussen lidwoord en zelfstandig naamwoord en voorspellingsbewustzijn (d.w.z. bewustzijn voor het nut van de lidwoorden bij het doen van de test). Alleen voorspellingsbewuste leerders bleken lidwoorden voorspellend te kunnen gebruiken en waren sneller. Hoewel geslachtsbewustzijn niet gerelateerd was aan voorspelling als zodanig, bleek dit...
wel een voorwaarde te zijn voor het ontwikkelen van voorspellingsbewustzijn. Men zou kunnen denken dat de voorspellende verwerking van lidwoorden het resultaat was van puur strategisch gedrag, maar dat lijkt onwaarschijnlijk. Ten eerste rapporteerden leerders voorspellingsbewustzijn als een gevoel dat ze tijdens de toets ontwikkelden en niet als een strategie die ze expres gebruikten. Ten tweede vereiste de voorspellende verwerking van lidwoorden goed geconsolideerde kennis van de verschillende combinaaties van lidwoorden en zelfstandige naamwoorden. Alleen het besef dat lidwoorden tijdens de toets gebruikt kunnen worden zou niet voldoende zijn geweest. Een meer waarschijnlijke verklaring is dat de leerders in de loop van de toets hun voorspellingsbewustzijn hebben ontwikkeld en dat dit bewustzijn het resultaat was van het opmerken van hun eigen voorspellend gedrag.

We vonden geen bewijs dat het werkgeheugen en het analytisch vermogen van de leerders tot een betere voorspellende verwerking van de lidwoorden leidden. De resultaten suggereerden echter wel dat het werkgeheugen daarbij indirect betrokken kan zijn, aangezien de neiging had te voorspellen of de leerders wel of niet bewust zouden worden van hun voorspellingen.

Voor de effecten van instructie waren de resultaten niet eenduidig. Leerders die het doelpatroon herhaalden en leerders die dat niet deden verschilden aanzienlijk in hun verwerkingspatronen. Toch vonden we geen duidelijk bewijs dat herhaling tot betere voorspelling op basis van lidwoorden leidde.

Dit proefschrift biedt een aantal methodologische inzichten die van grote waarde kunnen zijn voor toekomstig experimenteel tweedetaalverwervingsonderzoek. Ten eerste onderstrepen onze resultaten het belang en de waarde van het uitvoeren van grondige post-experimentele interviews met de deelnemers om de onderzoeksresultaten goed te kunnen interpreteren. Ten tweede verschaft het onderzoek een potentieel veelbelovend paradigma voor het meten van impliciete kennis op basis van items die leerders tijdens de blootstelling al hebben gehoord. Dit paradigma kan onderzoekers in staat stellen om subtiele signalen van impliciet patroonleren te detecteren die niet meetbaar zijn met behulp van grammaticaliteitsoordeeltaken met nieuwe items, omdat het gebruik van nieuwe items te veel verwarring bij de leerders kan creëren om impliciete kennis goed te kunnen meten. Ten slotte laat dit proefschrift zien dat de blootstellingsduur en het moment van toetsen van invloed kunnen zijn op de leerprocessen die worden gemeten. Om die reden moeten deze factoren zorgvuldig worden overwogen bij het ontwerpen van experimenteel onderzoek dat zich richt op een bepaald type leren.

Globaal gesproken biedt dit proefschrift een aantal nieuwe inzichten over de manier waarop kenmerken van taalaanbod en cognitieve capaciteiten van leerders betrokken kunnen zijn bij het leren van nieuwe taalpatronen. Het onderzoek laat zien dat de betrouwbaarheid van het taalaanbod de
leerresultaten niet alleen negatief, maar ook positief kan beïnvloeden, afhankelijk van hoe het leren verloopt. Het onderzoek biedt ook aanvullend bewijs voor de betrokkenheid van de werkgeheugencapaciteit en het analytisch vermogen bij het leren van tweedetaalpatronen. Bovendien zijn er aanwijzingen voor het bestaan van complexe interacties tussen cognitieve capaciteiten en instructie, waarbij de betrokkenheid van de cognitieve capaciteiten bij het leren afhankelijk kan zijn van de kenmerken van taalaanbod waaraan de leerders worden blootgesteld. Ten slotte bieden de resultaten enkele belangrijke inzichten met betrekking tot hoe verschillende niveaus van bewustzijn van de leerders van invloed kunnen zijn op het taalleren en de prestaties op zowel offline taken die de taalleeruitkomsten in kaart brengen als online taken die het proces meten.
Samenvatting
Maja Ćurčić was born in Novi Sad in Serbia, on 15 May 1988. In 2007 she finished a high school of economics (track tourism), after which she started a four-year bachelor program in the English language and literature at the University of Novi Sad. During her bachelor studies, she chose the linguistics track. She graduated cum laude in 2011, and decided to continue her studies abroad.

In 2011, she obtained the University of Groningen Talent Grant and started a two-year research master in linguistics at the University of Groningen. During this MA program she gained extensive experience with experimental research and graduated cum laude in 2013.

In September 2013 Maja started her PhD research (Explaining differences in adult second language learning: The role of language input characteristics and learners’ cognitive aptitudes project) at the University of Amsterdam, Amsterdam Center for Language and Cognition. During her PhD Maja was supervised by Sible Andringa and Folkert Kuiken, and she presented the results of her research at several conferences in the Netherlands and abroad (in France, Switzerland, Belgium, UK, and Finland). During her PhD, Maja also participated in teaching at Bachelor and MA level at the University of Amsterdam.

This dissertation summarizes the results of Maja’s four-year PhD research.