Foreign language learning in study-abroad and at-home contexts

Published by
LOT
Kloveniersburgwal 48
1012 CX Amsterdam
The Netherlands

phone: +31 20 525 2461

e-mail: lot@uva.nl
http://www.lotschool.nl

Cover illustration: Leilei Mao

ISBN: 978-94-6093-386-8
DOI: https://dx.medra.org/10.48273/LOT0601
NUR: 616

Foreign language learning in study-abroad and at-home contexts

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 11 oktober 2021
om 12.30 uur precies

door

Xiaoru Yu
geboren op 15 december 1989
te Henan (China)

Promotor:
Prof. dr. J.J.M. Schoonen

Copromotor:
Dr. E. Janse

Manuscriptcommissie:
Prof. dr. P.A.J.M. Coppen
Prof. dr. J.-M. Dewaele (Birkbeck, University of London, Verenigd Koninkrijk)
Prof. dr. J. Granfelt (Lunds universitet, Zweden)
Prof. dr. M.H. Verspoor (Rijksuniversiteit Groningen)
Dr. S.J. Andringa (Universiteit van Amsterdam)
Dr. E. Peters (Katholieke Universiteit Leuven, België)
Dr. M.B.P. Starren

This research was funded by a grant (NO. 201606090209) awarded by Chinese Scholarship Council (CSC) to Xiaoru Yu, and was co-financed by Centre for Language Studies (CLS) at Radboud University and Study Abroad Research from European Perspective (SAREP).

Table of Contents

Chapter 1: General introduction

"*If we think of language as a window through which we look at what the speaker is saying, in the case of first-language listening, the glass is very clean and we see through it without even noticing it is there; but in the case of second-language listening, the glass is dirty: we can see clearly through some parts, other parts are smudged, and yet other parts are so dirty we cannot see through them at all. We are very aware of the glass because it gets in the way.*" —Buck, 2001.

Buck's observation is a vivid analogy to illustrate the difficulties learners experience in L2 listening comprehension, which resonates in me ever since I read it. What is the nature of this dirty-glass problem? How does L2 proficiency affect the glass's transparency, and to what extent can we alleviate this problem by improving L2 learners' learning environment? How do we account for individual variability in second language acquisition? These questions motivated me to propose the current research project.

Studying abroad (SA) is often considered as the best L2 learning context (Freed, 1995), as it usually involves a language environment shift where learners have to inhibit the first language and immerse themselves in the target language (Jacobs, Fricke, & Kroll, 2016; Linck, Kroll, & Sunderman, 2009). Nowadays, with increasing global mobility (ignoring the 2019-2021 dip due to the COVID-19 pandemic), more and more students go abroad and find themselves in a new cultural and linguistic environment. According to Eurostat (2020), 1.3 million students from around the world came to Europe for tertiary education in 2018, 25% of whom were from Asia. China has been the largest source of outbound international students to Europe and worldwide for years, and is still developing international student mobility. The number of Chinese study-abroad students increased by 26.4% from 2014 (0.46 million) to 2018 (0.66 million) (Statista, 2020). Within Europe, the UK attracted the biggest influx of Chinese students, with the Chinese students accounting for 18.9% of the UK's foreign students in 2019/2020 (UK Parliament, 2021). The blooming international education highlights the need to further investigate how studying abroad influences students' language learning processes and outcomes. Furthermore, existing L2 research focused mainly on the effects of learning context on the development of speaking, reading, and writing, thereby largely ignoring learning context effects on listening development. The few studies that have investigated the relationship between learning context and L2 listening (e.g., Cubillos, Chieffo, & Fan, 2008; Llanes & Muñoz, 2009) often employed holistic measurement methods (e.g., spoken passage comprehension) to measure listening proficiency. Therefore, very little is known about the effect of learning contexts on finer (sub)components of listening comprehension.

This thesis examines how learning context and individual-difference factors may affect second language development, especially in listening comprehension. The rest of this section begins with a brief description of skill-based accounts (Anderson, 1983; DeKeyser, 2015) and usage-based accounts (Kemmer & Barlow, 2000; Langacker, 2000; Tomasello, 2009) of language acquisition. Through these theoretical lenses, we view second language acquisition as a process of acquiring knowledge (e.g., vocabulary and grammar) and improving processing skills (e.g., as in rapid word recognition), thus taking knowledge and processing efficiency as indicators of second language proficiency. That is, we will tackle the "dirty-glass" problem of second-language listening from the perspectives of knowledge and processing development. Furthermore, we will zoom in on spoken-language comprehension and elaborate upon our operationalization of L2 listening proficiency, based on existing theories of listening comprehension. Afterwards, a brief literature review of the effects of learning context and individual-difference factors on second language development will be presented. Finally, an overview of this thesis will be provided at the end of this section, which outlines our specific research questions and methodological approaches.

1.1 Skill-based accounts of second language acquisition

The quotation at the very beginning of this thesis accurately captured the effortfulness of L2 listening as compared to the apparent effortlessness with which most adults usually process their L1 (provided optimal listening conditions). It reflects a key concept in psychology, automaticity, which is generally used to refer to performing an activity with little cognitive effort (see Segalowitz, 2003). Anderson's (1983) Adaptive Control of Thought theory is one of the most convincing theories of cognitive skill acquisition. This theory makes a distinction between declarative knowledge, which can be described and is consciously held in memory, and procedural knowledge, which is much harder to describe and can be accessed or processed unconsciously. It holds that, at the early time of skill acquisition, performance largely relies on directed attention and deliberate control to employ declarative knowledge. With sufficient practice, skill acquisition gradually goes through a transition from a non-automatic stage to an automatic stage. The state of automaticity is realized when sequenced components of a skill become routinized or chunked, rendering the processes involved efficient and unavailable to conscious processing (Anderson, 1983).

Theories and empirical findings on general skill acquisition have led to applications and adaptations of these ideas to the field of second language acquisition. One of the most influential applications is DeKeyser's (2015) skill-based account of second language learning. DeKeyser (2007) argued that instructed learners start with explicit learning of declarative L2 knowledge, which gradually becomes proceduralized and eventually automatized through practice. DeKeyser (2015) claimed that the proceduralization of knowledge is not particularly time-consuming and may be complete after just a few instances. Automatization of knowledge, however, may take much longer as learners may need to go through a large amount of practice to decrease the processing time, error rates, and the amount of attention required when performing a linguistic task (DeKeyser, 2007). Segalowitz (2010) further proposed that processing stability, indexed by the coefficient of variation of response times (e.g., on a lexical decision task), is indicative of learners' language processing mechanisms. Segalowitz argued that changes in processing speed indicate quantitative changes (i.e., across-the-board speeding up), whereas changes in processing stability indicate qualitative changes (i.e., a restructuring of processes).

Under this theoretical framework, second language acquisition is both a matter of knowledge accumulation and of an increase in efficiency with which that knowledge can be processed (Hulstijn, Van Gelderen, & Schoonen, 2009). Hence our "dirty-glass" problem in L2 listening should be seen not only as a knowledge problem, but also as a processing problem. Accordingly, second language proficiency is operationalized as (vocabulary) knowledge and processing efficiency in this study. Building on DeKeyser (2015) and Segalowitz (2010), processing efficiency is further operationalized as a three-dimensional concept. In order for language processing to be efficient, processing must be accurate, fast and stable. Therefore, efficiency will be operationalized to entail accuracy, speed, and stability of processing.

1.2 Usage-based accounts of second language acquisition

Another set of theoretical lenses for understanding our "dirty-glass" problem come from usage-based theories of second language acquisition. A basic tenet of these theoretical accounts is that users' language ability emerges as a result of exposure to language input (Kemmer & Barlow, 2000; Langacker, 2000; Tomasello, 2009). Learners' L2 competence can, therefore, be assumed to be related to the quantity, quality, and type of L2 input they experience, which may affect both their language knowledge (e.g. vocabulary) and processing efficiency (e.g. speed of word recognition and grammar sensitivity). Usage-based theories also hold that language acquisition entails the "piecemeal learning" of constructions and the frequency-biased abstraction of grammatical patterns (Ellis, 2002). Ellis (2006) argued that L2 learners establish connections between specific linguistic forms and their meanings by frequently encountering and processing the target language input. Consequently, the strength of the connections, or mappings, between form and meaning is frequency-driven. Furthermore, usage-based theories distinguish token frequency, i.e., how often particular words or constructions occur in the input, and type frequency, i.e., how many distinct words or constructions there are in the input (see Ellis, 2002; Tomasello, 2000). Token frequency facilitates the entrenchment of linguistic items, rendering rapid language processing (Langacker, 1988; Tomasello, 2000), whereas type frequency promotes mastery of abstract patterns (Bybee & Hopper, 2001; Bybee & Thompson, 2000). Frequency effects in second language acquisition and processing have been empirically investigated through corpus-based analysis, i.e., counting and analysing statistical patterns of usage, and psycholinguistic methods, i.e., studying the influence of these statistical patterns on language processing (e.g., Bybee & Thompson, 2000; Crossley, Skalicky, Kyle, & Monteiro, 2019; Diependaele, Lemhöfer, & Brysbaert, 2013; Ellis, O'Donnell, & Römer, 2014; Garnier & Schmitt, 2016; Gollan, Montoya, Cera, & Sandoval, 2008; Kyle & Crossley, 2015).

According to the usage-based theories, L2 learners need to strengthen connections between forms and meanings, and to improve processing efficiency, through exposure to language input. However, in practice, foreign language learners may often only have access to impoverished and limited L2 input, in the form of textbooks and similar instructional materials. Such input almost always represents a distortion of naturalistic input in terms of distributional characteristics, most obviously word frequency. Textbooks for foreign language learning tend to cover a large number of word types at the cost of repetition of word tokens (see e.g., Matsuoka & Hirsh, 2010; Milton, 2009; Sun & Dang, 2020). For example, Matsuoka and Hirsh (2010) examined the repetition of words in an English language coursebook designed for learners at upper-intermediate proficiency levels. They found that West's (1953) second most frequent 1000 words were under-represented in the text. Only 603 out of those 1000 words actually appeared in the text, 33.3% of which occurred only once. At the same time, 1005 less frequent words were present in the coursebook, 66.4% of which also occurred only once. Due to the lack of repetition, foreign language learners might fail to learn words that have been covered in class, let alone that these words could become entrenched in the learners' language use. A related and perhaps even more severe problem for foreign-language learners concerns an overreliance on visually presented input (i.e., written input), which presents a stark contrast to the dominance of auditorily presented input in first language acquisition. Due to its transient nature, L2 learners often find it more difficult to process spoken than written discourse in the target language. It is not rare to see foreign-language learners who can read well but struggle while listening. The current research focused on L2 listening development in different learning contexts in order to follow up on these usage-based theories and on the above-mentioned limitations of L2 input in foreign-language learning contexts.

1.3 A componential view of listening comprehension

A number of theories have been put forward to account for the nature of listening comprehension. Goss' (1982) information processing model of listening categorized the process of listening into signal processing, literal processing and reflective processing. Signal processing involves segmenting the speech signal into potentially meaningful units. Literal processing is the initial assignment of basic meaning to the utterance. Inferences, if any, are restricted to basic understanding of literal meaning. Reflective processing is associated with critical or appreciative listening, where the listener evaluates the message and makes extensive inferences. Another view was proposed by Anderson (2015), initially brought up in 1995, that language comprehension consists of perception, parsing and utilization. The perception stage involves the encoding of the spoken message. Parsing is the process where the utterance is segmented and recombined according to syntactic and semantic cues to generate a mental representation of a combined meaning. Utilization is the final stage, in which the listener relates the mental representation of the sentence's meaning to existing knowledge, makes inferences and stores it in memory. Cutler and Clifton (1999) in their blueprint of the listener provided a detailed account of how a listener converts acoustic input into meaning, identifying four distinct yet interactive components: decoding, segmentation, recognition and integration. Listeners begin with the decoding process by separating a speech signal from background noise and transforming it into an abstract representation of phonological segments. Then the listener segments the continuous stream of speech into component parts by exploiting grammatical and semantic information as well as explicit segmentation cues (e.g., prosody). These segmentation cues are extracted from word recognition, utterance interpretation and early speech decoding processes. The recognition stage, including word recognition and utterance interpretation, overlaps with the segmentation process and is followed by the final stage where the listener integrates the utterance into its wider discourse context. The current thesis focuses on measuring L2 listeners' processing efficiency in word recognition, grammatical processing and semantic interpretation. These three target processes can be considered the subdivision of the "literal processing" in Goss' information processing model, the "parsing" stage in Anderson's language comprehension model or the "recognition" stage in Cutler and Clifton's blueprint of the listener.

The measurement of listening proficiency is one of the least understood and developed areas in language testing and assessment (Alderson & Banerjee, 2002; Batty, 2021). Speaking, reading and writing all involve explicit motor activities and/or manifest language output that can be used to uncover language processing to some extent. In contrast, the invisibility of the cognitive processes involved in listening makes it difficult to decompose the comprehension process. Hence L2 listening proficiency in the language-learning literature is usually assessed holistically, by means of spoken passage comprehension. Andringa, Olsthoorn, van Beuningen, Schoonen, and Hulstijn (2012) also explored listening comprehension of spoken passages (from the Dutch L2 State Exams). They attempted to predict listening performance by linguistic knowledge, as well as by the speed with which auditorily-presented linguistic information can be processed, and general cognitive ability. Results showed that linguistic knowledge (quantified by vocabulary knowledge and processing accuracy) and processing speed (e.g., in their grammatical processing and semantic processing tasks) could predict listening comprehension. Building on Andringa and colleagues' work, as well as on skills acquisition theories (Anderson, 1983; DeKeyser, 2015) and listening comprehension theories (Anderson, 2015; Cutler & Clifton, 1999; Goss, 1982) described above, we operationalize listening proficiency as auditory vocabulary knowledge on the one hand, and spoken-language processing efficiency on the other (i.e., accuracy, speed, and stability in word recognition, grammatical processing, and semantic processing). That is, the present thesis strives to probe into different components of listening comprehension in relation to the context in which language learning takes place and in relation to characteristics of the learner that may be associated with how easily individuals learn a foreign language.

1.4 Learning context: Study-abroad vs. at-home

Studying in a country where students' L2 is spoken as the native language guarantees abundant native L2 input, opportunities for L2 production, informative feedback, and interaction. Contrastively, even though nowadays it is relatively easy to have access to authentic L2 input through digital platforms (e.g., Netflix, Audible, and Apple News), at-home (AH) contexts may fall short because of relatively inadequate L2 exposure, over-reliance on rote learning, and limited opportunities for interactive communication with feedback. However, mixed findings on the relationship between studying abroad and second language learning have been reported in previous studies. Some found that SA learners had greater gains in knowledge of nativelike language usage (Foster, Bolibaugh, & Kotula, 2014), use of communication strategies (Lafford, 2004), grammar (Pliatsikas & Marinis, 2013b), accent (Martinsen, Alvord, & Tanner, 2014), pragmatic competence (Matsumura, 2001), writing proficiency (Sasaki, 2011), and oral proficiency (Segalowitz & Freed, 2004), but others reported marginal or no differences as a function of learning context in terms of grammar (Isabelli-García, 2010; Pliatsikas & Marinis, 2013a), and pragmatic comprehension (Taguchi, 2010). These mixed outcomes may relate to the fact that different studies focused on different aspects of language learning, which may not be equally sensitive to the potentially beneficial effects of the SA learning context.

As stated above, little is known on how studying abroad may affect the knowledge and processing aspects of L2 listening comprehension in particular. I therefore set out to evaluate the impact of studying abroad on L2 listening development in terms of auditory vocabulary knowledge and efficiency of the listening process. Note that it is usually difficult to observe significant progress in processing efficiency during a relatively short period of time in a normal foreign language learning condition, as it is well-documented that the development of L2 processing is slow and subtle (Lim & Godfroid, 2015). However, the dramatic environmental change entailed by a study abroad experience may accelerate the development of L2 processing and serve as a thrust strong enough to make changes in L2 processing more salient. In other words, by comparing second language development in study-abroad and at-home learning contexts, I will investigate the linguistic benefits and limits that stem from living in a country where the target language is spoken as the native language.

1.5 Individual differences and learning context

Individual differences abound in language acquisition. In the case of first language acquisition, fast children at 18 months can produce more than 250 words , while slow children produce fewer than 10 words at this stage (Fenson et al., 2006; for a review on individual differences in first language acquisition, see Kidd, Donnelly, & Christiansen, 2018). Second language acquisition has often been assumed to show great variability in acquisition rate, as well as in ultimate attainment (see Andringa & Dąbrowska, 2019; Dąbrowska, 2019). Previous research has shown that individual variability in second language acquisition can be explained, at least partially, by individual differences in cognitive, affective, social, and linguistic factors, such as language aptitude, working memory, motivation, socioeconomic status (for reviews of individual differences in second language acquisition, see e.g., Dewaele, 2009; Skehan, 2014).

However, second language learning contexts (e.g., regular classroom, study abroad, and domestic immersion) vary widely in terms of quality, quantity, and type of language exposure as well as opportunities for interaction, which inevitably contribute to the diversity of L2 development trajectories. Since learning context and individual-difference variables jointly affect the various aspects of L2 learning, understanding variability in L2 development requires joint investigation of these factors with careful consideration of their potential interplay (see Dörnyei, 2009; DeKeyser, 2012; Faretta-Stutenberg & Morgan-Short, 2018; Sanz, 2005). Indeed, some previous studies have suggested that individual differences in cognitive abilities interact with learning context in mediating second language learning processes and outcomes (e.g., Sunderman & Kroll, 2009; Tokowicz, Michael, & Kroll, 2004; Faretta-Stutenberg & Morgan-Short, 2018). Nevertheless, studies investigating the interaction between the effects of learning context and individual differences are quite limited in number and scope. Therefore, the present study will examine the effect of several individual-difference variables in various learning contexts to investigate the possible interplay between learning context, individual-difference variables, and L2 listening.

1.6 Thesis overview

The overall aim of this thesis is to investigate how environmental and individual-difference factors relate to second language development. More specifically, the following research questions will be addressed through three empirical studies of a study-abroad project and a systematic review of existing studies.

a) Do foreign language learners from three different learning contexts (i.e., AH-regular, SA-onset, AH-intensive) differ in their L2 listening proficiency? How do they compare to native listeners?
b) Do study-abroad learners show more improvement in listening proficiency than at-home learners over the course of an academic year?

c) What individual-difference factors are associated with listening proficiency and proficiency development across different learning contexts?

d) What is the overall effect of studying-abroad learning contexts, in comparison to at-home learning contexts, on second language development, based on results reported in existing study-abroad research?

The next four chapters (chapters 2 to 5) are devoted to answering questions a) to d), respectively. Chapters 2 to 4 are three empirical studies that involved the same participants, except that chapter 2 included an additional control group of native speakers of English. Chapter 5 is a systematic review which synthesized previous study-abroad research with a multi-level meta-analysis. Note that, as **Chapter 2** to **5** were written as stand-alone journal articles, they overlap in literature review and methodology in some degree.

Chapter 2 addresses the question of **whether foreign-language learners from different learning contexts differ in their L2 listening proficiency**. Three groups of Chinese postgraduates, together with a control group of native speakers were tested: non-English majors studying in China (AH-regular group), non-English majors newly arrived in the UK (SA-onset group), and English majors studying in China (AH-intensive group). Listening proficiency was measured with an auditory vocabulary size test (i.e., Peabody Picture Vocabulary Test Fourth Edition) and three language processing tasks (i.e., lexical access, grammatical processing, and semantic processing). The responses to the processing tasks were scored for accuracy, speed, and stability of processing. These three above-mentioned learning contexts differed mainly with regard to the amount of language exposure. Based on hours of instruction and self-directed learning activities, we assumed that the AH-intensive group had had higher EFL exposure than the AH-regular group, and that the SA-onset group would be somewhere in between the two domestic reference groups due to their preparation for studying abroad. According to usage-based theories of second language acquisition, we hypothesized that the AH-intensive group, which supposedly had the highest L2 exposure, would have largest auditory vocabulary and highest spoken-language processing efficiency, followed by the SA group and AH-intensive group in that order. We also expected that L2 learners with a larger auditory vocabulary size would have higher spoken-language processing efficiency. We were also interested in finding out whether the acquisition of vocabulary knowledge and that of processing efficiency would be equally sensitive to L2 learning context.

Chapter 3 investigates **the effect of studying abroad on L2 listening development**. As in Chapter 2, we focused on listening proficiency in terms of knowledge and processing measures. We hypothesized that study-abroad learners would make more progress than their domestic counterparts in terms of both auditory vocabulary size and spoken-language processing efficiency. To test this hypothesis, we invited Chinese international non-English-major postgraduates studying in the UK (SA group, with no previous study-abroad experience), Chinese domestic English-major postgraduates (AH-intensive group) and domestic non-English-major postgraduates (AH-regular group) to participate in a battery of listening tests at the beginning of their postgraduate program and again after one academic year. Note that **Chapter 3** was built upon **Chapter 2** whose participants were invited back after one academic year to take a post-test.

Chapter 4 investigates **how individual-difference factors and learning context relate to L2 listening development**. L2 learners may acquire language at different paces due to individual differences in cognitive, emotional, social, and linguistic factors (see e.g., Dewaele, 2009; Dörnyei, 2005; Robinson, 2002; Skehan, 2014), even when they are learning a language in exactly the same environment. Language aptitude, working memory, mental well-being, language exposure, and social interaction are individual-difference factors that have been associated with language learning (see e.g., Dąbrowska, 2019; Dewaele, 2009; Gass & Mackey, 2007; Granena, 2013; MacIntyre, Gregersen, & Mercer, 2019; Robinson, 2005; Skehan, 2014), and it is conceivable that some of them may be more relevant in certain learning contexts than in others. This chapter aimed to unravel the complex relationships between language learning and individual-difference factors in different learning contexts. We used a pre- and post-test design, testing Chinese learners of English at the beginning and end of an academic year, to examine the effect of several individual-difference variables in three learning contexts and to investigate possible interaction effects between learning context and individual-difference variables on L2 listening development. This chapter was built upon the previous two chapters. After the participants finished the pre- and post-tests described in **Chapter 3**, they were asked to take a series of tests and questionnaires to measure individual-difference factors, i.e., language aptitude, working memory, mental well-being, language exposure, and social interaction.

Chapter 5 is a systematic review paper that evaluates **the impact of studying abroad on L2 development based on existing research**. This chapter quantitatively synthesized twenty studies, all having a between-group pretest-posttest design (i.e., comparing second language development in study-abroad and at-home learning contexts over a certain period of time). These studies investigated the effect of studying abroad on various aspects of second language development. I fitted multilevel meta-analytical models to estimate the overall effect size of studying abroad and to explore possible moderating factors of studying-abroad effects, such as length of stay, evaluated outcome measures, and cultural distance between the original and the hosting countries.

Finally, **Chapter 6** summarizes the findings from previous chapters. A general discussion of the results is presented within a wider context. In this chapter, methodological recommendations and possible directions for future research will be provided.

Chapter 2: Breaking down listening comprehension: Vocabulary knowledge and processing efficiency in English-as-a-foreign-language (EFL) learners

Abstract

This study examines L2 listening proficiency of English-as-a-foreign-language (EFL) learners from three learning contexts. Based on DeKeyser's (2015) skill-based account of second language acquisition, we view L2 development as a process of knowledge accumulation and processing automatization, thus taking knowledge and processing efficiency as indicators of listening proficiency. 165 Chinese postgraduates, together with 23 native English speakers (control group), were administered a battery of listening tasks: 60 non-English-major students in China (at-home-regular group), 53 non-English-major students newly arrived in the UK (study-abroad-onset group), and 52 English-major students in China (at-home-intensive group). The amount of EFL exposure these three nonnative groups had received was lowest for the at-home-regular group, higher for the study-abroad-onset group, and highest for the at-home-intensive group. The listening tasks measured auditory vocabulary knowledge and spoken-language processing efficiency (i.e., accuracy, speed and stability of L2 speech processing) in word recognition, grammatical processing, and semantic processing. The results showed that the at-home-intensive group had a larger vocabulary size than the study-abroad-onset group, who in turn outperformed the at-home-regular group. The at-home-intensive and study-abroad-onset groups did not differ in any of the processing measures, but they both outperformed the at-home-regular group in accuracy and speed of processing across the processing tasks. That is, increasing amount of EFL exposure was associated with larger vocabulary size, but not always with higher processing efficiency. This study suggests knowledge accumulation and processing automatization may not be equally affected by EFL exposure.

2.1 Introduction

Second language acquisition (SLA) is both a matter of knowledge accumulation and of an increase in the efficiency with which that knowledge can be processed (DeKeyser, 2015; Hulstijn, Van Gelderen, & Schoonen, 2009). In English as a Foreign Language (EFL) learning contexts, only limited L2 exposure is available for learners to utilize for accumulating language knowledge and improving processing efficiency. Different EFL learning contexts may offer different amounts of exposure, which can be assumed to be related to learners' second language competence in terms of both knowledge and processing efficiency. Studies thus far have only provided evidence that L2 exposure modulates knowledge accumulation, such as the acquisition of vocabulary, collocations and syntactic constructions (Dąbrowska & Street, 2006; Milton, 2009; Sonbul & Schmitt, 2013; Toomer & Elgort, 2019). Despite the fact that L2 processing mechanisms have been investigated extensively in the areas of lexicon, morpho-syntax, phonology and bilingualism (Jiang, 2018; Pütz & Sicola, 2010), it is generally unclear whether and to what extent automatization of L2 processing is modulated by L2 exposure. This study aims to investigate knowledge and processing aspects of L2 listening proficiency in EFL learners from three learning contexts that supposedly differ with regard to amount of L2 exposure.

Out of the four language skills, listening is the least researched area in the field of SLA (Buck 2001; Vandergrift, 2007). Perhaps relatedly, the assessment of listening abilities is one of the least understood and developed areas in language testing and assessment (Alderson & Banerjee, 2002). The invisibility of the cognitive processes involved in listening makes it difficult to decompose the comprehension process. L2 listening tests are usually confined to holistic measurement methods like spoken passage comprehension. Scores thereof are used as a rough approximation of listening proficiency, but reflect little about the cognitive underpinnings of the listening process. The process-oriented listening studies (i.e., Farrell & Mallard, 2006; Goh, 2000) in SLA research often resort to retrospective techniques (i.e., questionnaires, stimulated recall and interviews), introspective techniques (i.e., thinking aloud), and observation (i.e., analyzing video recording of conversational interaction) to understand the L2 listening process (Vandergrift, 2007). However, these approaches are subject to subjectivity coming from either participants or raters, as they rely heavily on participants' awareness of the listening process or raters' perception of conversational difficulties. Alternatively, drawing upon methodologies of first language processing in psycholinguistics, a growing body of research adopts an experimental approach to study nonnative listening processing. Test stimuli are manipulated to allow for detection of the subtle nuances in language processing that are not easily identifiable by existing standardized tests. This psycholinguistic research strand usually focuses on factors determining how easily certain linguistic structures can be acquired, or on cognitive limitations of nonnative processing (Jiang, 2018). However, how L2 learners develop language processing efficiency with respect to environmental factors has received relatively little attention. This study is devoted to breaking down L2 listening comprehension into smaller components and systematically examining them with specifically-developed listening materials in an empirical testing environment. Listening proficiency is operationalized as auditory vocabulary knowledge and processing efficiency of three listening subprocesses (i.e., lexical access, morphosyntactic processing, and semantic processing). In other words, we take a componential view of listening comprehension to study L2 auditory vocabulary and spoken-language processing efficiency in EFL learners from different learning contexts.

2.1.1 Listening comprehension processes

A number of theories have been put forward to account for the nature of listening comprehension. Goss' (1982) information processing model of listening categorized the process of listening into signal processing, literal processing, and reflective processing. Signal processing involves segmenting the speech signal into potentially meaningful units. Literal processing is the initial assignment of basic meaning to the utterance. Inferences, if any, are restricted to basic understanding of literal meaning. Reflective processing is associated with critical or appreciative listening, where the listener evaluates the message and makes extensive inferences. Another theory was proposed by Anderson (2015), initially brought up in 1995, that language comprehension consists of perception, parsing, and utilization. The perception stage involves the encoding of the spoken message. Parsing is the process where the utterance is segmented and recombined according to syntactic and semantic cues to generate a mental representation of a combined meaning. Utilization is the final stage, in which the listener relates the mental representation of the sentence's meaning to existing knowledge, makes inferences and stores it in memory. To provide a detailed account of how a listener converts acoustic input into meaning, Cutler and Clifton (1999) proposed a blueprint of the listener identifying four distinct yet interactive components: decoding, segmentation, recognition, and integration. Listeners begin with the decoding process by separating a speech signal from background noise and transforming it into an abstract representation of phonetic segments. Then the listener segments the continuous stream of speech into component parts by exploiting grammatical and semantic information as well as explicit segmentation cues (e.g., prosody). These segmentation cues are extracted from word recognition, utterance interpretation and early speech decoding processes. The recognition stage, including word recognition and utterance interpretation, overlaps with the segmenting process and is followed by the end stage where the listener integrates the utterance into its wider discourse context. The present study measures L2 listeners' processing efficiency in word recognition, grammatical parsing, and semantic interpretation. The three targeted processes can be considered the subdivisions of the "literal processing" in Goss' information processing model, the "parsing" stage in Anderson's language comprehension model or the "recognition" stage in Cutler and Clifton's blueprint of the listener.

Firstly, word recognition presumably becomes more efficient along with increasing exposure. In both native and nonnative processing, when a word is heard, not only the word itself but also a set of words sharing its phonetic, semantic or contextual features are activated (see Weber & Scharenborg, 2012). It is important that the language user can map the auditory representation to the correct lexical item while activating the accurate semantic representations. More proficient language users are supposed to be better able to select the target item from its competitors. Secondly, grammatical information is used in constructing an initial analysis of an utterance. Both global grammatical information (e.g. about possible phrase structure configurations) and plausibility information entailed in individual lexical items (e.g. the possible argument structure assigned to a verb) contribute to the construction of the grammatical structure of a sentence. Morphology, syntax and function words display structural features of a language and are areas of major interlanguage differences. Morphosyntax is hard to be internalized and errors pervade even in advanced L2 learners' speech (see DeKeyser, 2005). Adult L2 learners rely more on lexical, semantic, and pragmatic information than grammar in sentence comprehension, and effects of syntactic structure that were seen in native speakers appear to be absent in L2 processing (Clahsen & Felser, 2006). Thirdly, semantic interpretation of a sentence is an incremental process that keeps up with the words as they are heard (Cutler & Clifton, 1999). Listeners instantly relate the propositional representation of an utterance to their world knowledge activated during word recognition. This means that proficient learners can quickly notice semantic violations in implausible sentences (e.g., "Most bicycles have seven wheels"). Due to weak language proficiency, it is relatively difficult for L2 learners to notice any potential discrepancy between their world knowledge and an L2 sentence, particularly when under time pressure as is usually the case in nonnative listening.

2.1.2 Operationalization of processing efficiency

Automaticity in language processing implies that language tasks are performed with little cognitive effort (Segalowitz, 2003). Shiffrin and Schneider (1984) defined automaticity as "a fast, parallel, fairly effortless process that is not limited by short-term memory (STM) capacity, is not under direct subject control, and is responsible for the performance of well-developed skill behaviors". The state of automaticity is realized when sequenced components of a skill become routinized or chunked, rendering the processes involved efficient and unavailable to conscious awareness (Anderson, 1983). DeKeyser (2007) argued that second language acquisition, like any other type of skill acquisition, experiences a transition from a non-automatic stage to an automatic stage. According to DeKeyser's (2015) skill-based account of second language acquisition, instructed learners start with explicit learning of declarative L2 knowledge, which gradually becomes proceduralized and eventually automatized through practice. The proceduralization of knowledge is not particularly time-consuming and may be complete after just a few trials/instances. Automatization of knowledge, however, may take much longer as learners may need to go through a large amount of practice to decrease the reaction time, error rates, and the amount of attention required when performing a linguistic task.

Researchers describe automaticity in different ways, e.g., in terms of fast, ballistic, load-independent, effortless, and unconscious processing (for a review, see Segalowitz, 2003). Segalowitz pointed out that these definitions refer to logically distinct possibilities in the way psychological mechanisms may operate, and that using "automaticity" as a short-hand term incurs confusion and imprecision. Therefore, despite the existing line of L2 automaticity research, the present study shifts from the theoretically charged term "automaticity" to the more neutral term "processing efficiency". By using the gradient concept of processing efficiency, we also acknowledge the fact that it is still questionable whether second language processing, even for highly proficient L2 learners, could be truly automatic or not (Dijkgraaf, Hartsuiker, & Duyck, 2019). Furthermore, building on Segalowitz's (2010) proposition to take processing speed, processing stability and processing flexibility as key properties that determine the degree of automaticity, this study operationalizes processing efficiency as processing speed, processing stability and processing accuracy. According to Segalowitz (2010), faster processing speed and greater processing stability are indicative of different changes in learners' underlying processing mechanisms. Whereas faster processing speed indicates *quantitative* changes (i.e., across-the-board speeding up), greater processing stability is argued to indicate *qualitative* changes (i.e., a restructuring of processes).

2.1.3 The role of language exposure in second language acquisition

A basic tenet of usage-based theories of language acquisition is that users' language ability emerges as a result of language exposure (Kemmer & Barlow, 2000; Langacker, 2000; Tomasello, 2003). Ellis (2006) argues that L2 learners establish connections between specific linguistic forms and their meanings by frequently encountering and processing the target language input. Consequently, the strength of the connections (or mappings) between form and meaning is frequency-driven. Token frequency facilitates the entrenchment of linguistic units, rendering rapid language processing (Langacker, 1988; Tomasello, 2000), whereas type frequency promotes mastery of abstract patterns (Bybee & Hopper, 2001; Bybee & Thompson, 2000; N. Ellis, 2002). Empirical studies have shown positive correlations between mastery of words and grammatical constructions and how often these words or constructions occur in language input. This has been shown through various methods like corpus-based analysis and classroom field research (e.g., Ellis, O'Donnell, & Römer, 2014; McDonough & Kim, 2009). On a related note, some researchers (e.g., Lee, 2007; Muranoi, 2000; Robinson, 1997) manipulate input distribution in order to make certain aspects of language salient. However, studies so far have yielded mixed findings on the advantages of using enhanced input against unstructured exposure in facilitating SLA (for a review, see R. Ellis, 2016; Han, Park, & Combs, 2008).

Furthermore, some studies speculate on a stabilization or fossilization phenomenon where L2 learners cease to improve their language proficiency regardless of further exposure to target language (N. Ellis, 2006; Han, 2013; Han & Odlin, 2006; Selinker, 1972). This phenomenon suggests a nonlinear relationship between exposure and language acquisition. Note, however, that most of these studies have been carried out in English as a second language (ESL) contexts where learners are immersed in their L2. It is not clear whether learners could already reach this point of stabilization with little exposure as in EFL contexts or only in contexts where they are exposed to abundant naturalistic experience. Given the perceived differences between language learning in an ESL context and an English as a foreign language (EFL) context (Freed, 1995), the research findings on the role of language exposure in ESL contexts may have limited validity for language learning in EFL contexts. One of the characteristics that may differentiate EFL and ESL learners is the efficiency or level of automaticity with which they can use their additional language.

2.1.4 The current study

The present study aims to investigate whether and how EFL learners from three different learning contexts differ in knowledge and processing aspects of L2 listening comprehension. We expect learning contexts associated with higher exposure amounts to be associated with larger vocabulary and higher language processing efficiency. We are also interested in finding out whether the acquisition of vocabulary knowledge and that of processing efficiency are be equally sensitive to L2 learning context.

The different learning contexts are represented by three groups of Chinese EFL students who were just enrolled in master programs when participating in this research: (a) an at-home-regular (AH-regular) group, consisting of domestic students taking non-English majors, (b) a study-abroad-onset (SA-onset) group, consisting of newly-arrived Chinese international students in the UK who majored in any subject but English language and culture and had no previous study-abroad or domestic-immersion experience, and (c) an at-home-intensive (AH-intensive) group, consisting of domestic students majoring in English language and culture. They took our tests measuring L2 vocabulary knowledge and processing efficiency, together with a group of English native speakers whose performance was taken as a reference point. As the Chinese Education Ministry specifies that English language courses should account for ten percent of non-English major university students' curriculum, we reasoned that non-English major students on average receive less English-language input than English-major students. Furthermore, we reasoned that non-English majors who had learned English intensively in preparation for studying abroad have in general had more English exposure than those who had not.

Therefore, we expected that the AH-intensive group would have higher EFL exposure than the AH-regular group and that the SA-onset group would be in between (see section 2.2.1 for more information).

We firstly compared native to nonnative performance to reveal the distance between nonnative speakers' current performance and ceiling (i.e., native) performance. Comparing L2 learners against native speakers helps determine what aspects of the language system have or have not been fully acquired (Montrul, Foote, & Perninan, 2008). The three nonnative groups were then compared against each other in order to study second language acquisition in different EFL learning contexts. The following research questions were addressed in this study:

1. Do native participants and nonnative participants differ in listening proficiency as quantified by vocabulary knowledge and language processing efficiency?
2. Do EFL learners from the three different learning contexts differ in listening proficiency as quantified by vocabulary knowledge and language processing efficiency?

2.2 Methodology

2.2.1 Participants

188 university-level students, consisting of three EFL groups and one native English group, participated in this study: (a) 60 Chinese domestic non-English-major postgraduates (the AH-regular group); (b) 53 Chinese international postgraduates who just arrived in the UK with no previous study-abroad or domestic immersion experience (the SA-onset group); (c) 52 Chinese domestic English-major postgraduates (the AH-intensive group), (d) and 23 native speakers of English attending universities in the UK.

The three EFL samples in this study are representatives of the most typical English learning contexts in China's higher education system. They had all learned English for at least ten years (i.e., 3 years of middle school, 3 years of high school, 4 years of bachelor program) in China. The AH-intensive group were English majors whose courses were mainly given in English (around 1620 hours in total as prescribed by their bachelor programs), including comprehensive English, speaking, listening, reading, writing, literature, culture, and linguistics. In contrast, the AH-regular and the SA-onset groups were non-English majors who had to spend most of their time on another major. They only had one English class once per week (around 144 hours in total as prescribed by bachelor program), which was mostly reading-oriented to prepare them for the written exams they had to take during college. Hence the AH-intensive group had more exposure to the English language than the AH-regular and the SA-onset group. Note that the exposure difference between English majors and non-English majors should be larger than the difference in hours of instruction reported previously, as time investment for before-class preparation and after-class home assignments were not counted in. Furthermore, although the AH-regular and the SA-onset group had the same amount of English classes, the latter had to spend considerably more time learning English in preparation for studying abroad. Consequently, the exposure level of the SA-onset group was assumed to be higher than that of the AH-regular group but lower than that of the AH-intensive group.

All participants were aged between 18-28 years old to minimize the potential effect of age, an extra-linguistic factor, on time-sensitive measures. Students in the AH-regular learning context who scored lower than 500 (out of 710) in College English Test Band 6 (CET-6) turned out to have very low performance in the sentence comprehension tasks in our pilot study. As our focus was on relatively automatic processing, which can only be investigated sensibly in learners with at least intermediate language proficiency, students in the AH-regular learning context had to score above 500 (CET-6 scores form a normal distribution with 500 as the mean score) in order to participate. To minimize additional noise caused by different arrival times, we only invited newly-arrived students who had not entered the UK more than one month prior to taking our tests.

Furthermore, a one-way ANOVA was conducted to compare the CET-6 scores of the three nonnative groups (see Table 2.1). Results show that there was a significant difference in mean scores between the groups (*F*(2,144)=7.964, *p* = .0005, η^2 = .102). Post-hoc tests show that the AH-regular and the SA-onset group did not differ significantly (*p* = .221), while the AH-intensive group significantly outperformed the AH-regular (*p* = .03) and the SA-onset (*p* = .0003) group. Nevertheless, these results have to be interpreted with caution because participants took the CET-6 test sometime during their bachelor program, which could have been a considerable amount of time (up to three years) before participating in this study.

Table 2.1: Background information of participants

Group	Sample size	Gender (F, M)	Mean age (*SD*)*	Mandatory English credits (Bachelor)*	Mean CET-6 score (SD)*
AH-regular	60	30, 30	22.5 (1.0)	8 credits	542 (32.6)
SA-onset	53	47, 6	22.4 (1.0)	8 credits	528 (54.7)
AH-intensive	52	48, 4	22.5 (1.0)	> 90 credits	563 (38.8)

Native	23	13, 10	22.4 (2.7)	Not applicable	Not applicable

**Notes: 1. Not all nonnative participants reported their age and CET-6 score. Therefore, means and SDs for these two variables are based on smaller sample sizes (i.e., 143 and 147 out of 165 nonnatives for Age and CET-6 score, respectively).*

2. One credit equals around 18 hours of English classroom instruction (and a corresponding amount of homework assignments).

2.2.2 Instruments

All testing instruments, i.e., a lexical access task, a grammatical processing task, a semantic processing task, and a Peabody Picture Vocabulary Test (PPVT), were computerized with *Presentation* software. The use of experimental software allowed for accurate stimulus presentation and time measurement. The first three tasks measure language processing efficiency in lexical access, syntactic parsing and semantic proposition formation, as will be elaborated subsequently. Accuracy rate, reaction time (RT) and coefficient of variation (CV) were recorded as measures of processing efficiency ($CV = SD_{RT}/Mean_{RT}$; for more information about CV see Segalowitz, 2010). The PPVT was used to measure auditory receptive vocabulary size, which was used as an indicator of declarative knowledge.

Lexical access task

The lexical access task was used to simulate various kinds of lexical access in controlled conditions where auditory and visual stimuli were simultaneously presented. Six training trials preceded 60 experimental trials. If no response was given within 4s, the current trial would end and the experiment would proceed to the next trial automatically. On each trial participants heard a word and saw a picture at the same time, and were asked to make a speedy judgment whether the spoken word and the picture matched by pressing either a “Yes” or “No” button on a button box. Whenever a word and a picture did not match, they could be classified into phonetically similar pairs, meaning-related pairs and random pairs. In the first category, the word participants heard and the depicted word, e.g., “kite” and “cat”, formed a phonetically similar pair. These test items measured lexical access with the visual presence of a phonetically similar target word. In the case of meaning-related pair, the spoken word and the picture were from the same semantic category, but did not overlap phonetically, i.e., “orange” and “apple”. This type of test items measured lexical access with the visual presence of a semantically similar word. A *random pair* contains an unrelated pair of words, i.e., “frog” and “doctor”. The last category measures lexical access without involving a phonetically or semantically similar candidate as in the first two categories.

Grammatical processing task

The grammatical processing task was designed to focus on how sensitive participants were to the most basic grammatical structures (e.g., plural "-s", 3rd person singular "-s", tense) rather than to measure the depth of their grammar knowledge. This task contains six training trials and sixty experimental trials. The testing materials, partially originating from Kersten (2010), Waters, Caplan, & Rochon (1995) and Weist (2002), have been manipulated purposefully to tap into grammatical processing. In this task, participants listened to a sentence and saw two pictures on the screen simultaneously, and were required to quickly indicate which picture matched the sentence they had heard by pressing either the "Left" or "Right" button on a button box. The maximum reaction time for each stimulus was set to 8 seconds for this task. To give a correct response, participants needed to catch the grammatical cue in the sentence they had just heard. These cues could be put into three categories: morphological cues, syntactic cues and function words (see Appendix C for an example stimulus of each category). The morphological cues included plural -s, 3rd person singular –s, tense and aspect. For example, participants heard the sentence "The woman puts her shirts on the desk" and saw two pictures. There was a woman putting one shirt on a desk in the first picture and three shirts on a desk in the second picture. The plural "-s" in the sentence was the morphological cue leading to the correct picture. The syntactic cues contained dative, passive and cleft constructions, and relative clauses. For example, participants heard the sentence "It is the horse that the pig pushes" and saw two pictures (one with a horse pushing a pig, while in the other one there was a pig pushing a horse). The cleft sentence structure was the syntactic cue indicating the correct picture. The function-word cue contained prepositions and conjunctions. For example, participants heard the sentence "The truck is standing across the railroad tracks" and saw two pictures: In the first one the truck was standing across the railroad tracks, and in the second one it was standing along the railroad tracks. The preposition "across" in the sentence was the function-word cue in this case. The first two categories (forty trials) were taken as target stimuli measuring morphosyntactic processing. The last category (twenty trials) was treated as filler items and excluded from the analysis as it did not fully qualify to indicate grammatical/morphosyntactic processing.

Semantic processing task

This instrument is a dichotomous plausibility judgment task to measure how fast participants can form a semantic proposition when listening to a sentence. Participants were asked to decide as accurately and rapidly as possible whether a sentence made sense or not. Nonsensical sentences violate factual knowledge (e.g., “A horse is an animal that can fly”) or logic (e.g., “If you eat too much, you can get too thin”). Six practice trials preceded fifty experimental trials. A response had to be given within eight seconds. The test items were adapted from Lim and Godfroid (2015). A sentence should be considered implausible when its meaning disagrees with listener’s prior knowledge. Listening comprehension is measured by determining if listeners can successfully make plausibility judgments of the sentences they have heard.

Vocabulary size test

The Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4, Dunn & Dunn, 2007) was used to measure auditory receptive vocabulary size. Participants saw four numbered pictures on the screen and were presented with an auditory word stimulus. Then they needed to press a numbered button corresponding to the picture that best illustrates the meaning of the word they had just heard. For example, the participants heard “pick” and saw four pictures: the first one is about peeling an apple, the second about pouring coffee, the third about picking oranges, and the fourth about cracking an egg. The third picture is the correct answer in this case. There is no time pressure for participants to give responses and each word can be played multiple times if needed.

The PPVT is designed to measure receptive (auditory) vocabulary size of English native speakers aged from 2:6 to over 90 years. It consists of 228 test items grouped into 19 sets which are arranged in order of increasing difficulty in the test. The test’s vocabulary items are not restricted or biased to the vocabulary of any specific purpose or discipline. When participants made more than eight errors out of twelve stimuli in one set, the test would end. Afterwards, an accuracy-based raw score was calculated as the measure of vocabulary size, which is taken as a proxy for declarative lexical knowledge accumulated in SLA.

Other instruments

Apart from the previously-mentioned testing instruments, a background questionnaire and an auditory processing speed test were used to screen participants and control for potentially confounding factors. The background questionnaire collects information about participants’ age, language learning experience, and previous standardized test scores. The auditory processing speed test measures non-linguistic individual differences in auditory processing. Participants were asked to indicate as fast as possible whether a tone was high (400 Hz) or low (300 Hz) in this task. Note that the participant groups did not differ significantly in either accuracy or RT: mean accuracy rates (SDs between brackets) were 0.98 (0.03), 0.98 (0.03), 0.99 (0.01) and 0.99 (0.01) for the AH-regular, SA-onset, AH-intensive, and native group, respectively; mean RTs were 456 (90), 464 (113), 494 (96) and 482 (120).

2.2.3 Data collection

Background questionnaires and recruitment advertisements were sent out beforehand to university students in China and the UK. After screening questionnaires, the researcher invited eligible participants to take the language tests. After a detailed description of all the tasks, the researcher asked participants to sign an informed consent form before taking the tests. Afterwards, participants received financial compensation for their participation. The tasks were administered in a fixed order. Participants' responses (e.g. reaction time and accuracy) in each task were automatically logged by the *Presentation* software for later analysis.

2.2.4 Data cleaning

The purpose of data cleaning was to identify and exclude deviant items and participants. Data of the reference group of native speakers were used to identify ambiguous items. Items with lower than 80% accuracy rate for native speakers were regarded as ambiguous items and were therefore excluded. As a result, we excluded four items out of forty in the grammatical processing task and three items out of fifty in the semantic processing task. Furthermore, participants with exceptionally low accuracy scores were considered to be performing very poorly, possibly due to inattentiveness. Participants who scored more than 3 standard deviations below the group mean in the vocabulary test were excluded from this study. Consequently, one participant in the AH-intensive group was excluded. No participants were excluded for low accuracy in any speeded test as all of them scored 50% or higher. In addition, reaction times (RTs) of incorrect responses in each experiment were excluded from the RT analysis. Considering the time needed for a valid button press response, reaction times below 250 ms (measured from audio onset) were considered invalid responses and were removed as well. Hence, accuracy measures were calculated across all valid trials, while RTs were calculated only on the basis of correct responses.

2.2.5 Statistical analysis

We analyzed the measures discussed previously via regression modeling in R version 3.5.1 (R Core Team, 2018). The *glmer*, *lmer* and *lm* functions in the package *lme4* (Bates, Maechler, Bolker, & Walker, 2015) were used to fit logistic and linear regression models, the optimizer being bobyqa. P values were calculated and added into model outputs with the package *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2017).

Two sets of models were fitted: the first one to compare first between native and nonnative performance and the second one to compare the performance among nonnative groups. The model selection procedures adhere to the principle of applying the maximal random effects structure on the premise of model convergence. The predictors in all the models are theoretically motivated, making model stripping unnecessary. As accuracy is a binary variable (coded as either 0 or 1), logistic regression models were fitted to predict response accuracy in the language processing tasks. Accuracy models took group, task and their interaction as fixed-effects predictors, and included maximal by-participant and by-item random intercepts and slopes. As RT is a continuous variable, linear mixed-effected regression models were fitted to predict response speed in the language processing tasks. The distribution of RTs deviated from a normal distribution with a long right tail, so log transformation was conducted to normalize the RT distribution. Since the audio duration of test items and trial number might also affect participants' response speed, these factors were entered in the RT models as fixed-effect variables, together with the whole set of variables contained in accuracy models. Moreover, inclusion of auditory speed (tone classification) as a covariate in the RT models did not change the research findings: i.e., the patterns of significant effects remained the same. Hence, for practical reasons, we present the results of the models without auditory speed. As for the processing stability measure, CV is an aggregated measure calculated on task level for each participant. The CV models had group, task and their interaction as fixed-effects predictors and a by-participant random intercept, without any item-level variables. For all the models of processing efficiency measures, the categorical variables, i.e., group and task, were coded using simple coding scheme which provides the canonical main effects by specifying the intercept as the grand mean, instead of the mean of the reference level. As for vocabulary size, linear regression models, with group as a fixed effect and no random effects, were fitted to compare group differences in vocabulary knowledge.

2.3 Results

Descriptive statistics for the performance of the participant groups on the language tasks are displayed in Table 2.2. As expected, the accuracy level of the native group almost reached the ceiling level across all tasks. The differences between native and nonnative groups seemed to be larger than the differences among the nonnative groups across all measures. This section will give a detailed description of the statistical results (for model estimates, see Appendix A; for data visualization, see Appendix B).

Table 2.2: Descriptive statistics of task performance in the language tests

Task	Measure	Group	*Mean*	*SD*
Lexical access task	Accuracy	Nonnative	.86	.06

		Native	.97	.03
		AH-regular	.83	.06
		SA-onset	.88	.04
		AH-intensive	.87	.05
	RT	Nonnative	1200	263
		Native	759	90
		AH-regular	1284	288
		SA-onset	1136	248
		AH-intensive	1168	226
	CV	Nonnative	.34	.09
		Native	.23	.07
		AH-regular	.35	.08
		SA-onset	.33	.08
		AH-intensive	.34	.09
Grammatical processing task	Accuracy	Nonnative	.85	.07
		Native	.96	.03
		AH-regular	.83	.08
		SA-onset	.86	.07
		AH-intensive	.86	.06
	RT	Nonnative	3300	438
		Native	2471	336
		AH-regular	3379	453
		SA-onset	3281	430
		AH-intensive	3228	423
	CV	Nonnative	.38	.05
		Native	.38	.04
		AH-regular	.38	.05
		SA-onset	.38	.05
		AH-intensive	.39	.06
Semantic processing task	Accuracy	Nonnative	.90	.10
		Native	.98	.02
		AH-regular	.84	.12
		SA-onset	.93	.07
		AH-intensive	.93	.05
	RT	Nonnative	3237	583
		Native	2374	191
		AH-regular	3598	670
		SA-onset	3051	449
		AH-intensive	3004	347
	CV	Nonnative	.28	.06
		Native	.20	.07
		AH-regular	.29	.06

		SA-onset	.26	.05
		AH-intensive	.28	.06
Vocabulary size test	Score	Nonnative	130	27
		Native	205	7
		AH-regular	115	29
		SA-onset	131	23
		AH-intensive	147	17

Note: Sample sizes of the nonnative, native, AH-regular, SA-onset and AH-intensive groups are 164, 23, 60, 53 and 51 respectively.

2.3.1 Comparisons between nonnative and native groups

Processing accuracy

The processing accuracy model showed significant task effects (i.e., there was a significant difference between the lexical access and the grammatical processing task, but not between the lexical access and the semantic processing task; see Appendix Table A1). Native participants were more accurate than participants in the nonnative group across the three processing tasks (β = -1.54, *SE* = 0.22, z = -7.17, $p < .001$). The difference between native and nonnative groups in the lexical access task was not significantly different from that of the grammatical processing task (β = 0.13, *SE* = 0.36, z = 0.37, $p > .05$) or that of the semantic processing task (β = -0.47, *SE* = 0.44, z = -1.08, $p > .05$). So there was no significant interaction between group and task variables.

Processing speed

The processing speed model showed significant task effects (see Appendix Table A2). Similar to processing accuracy, the difference between native and nonnative groups in processing speed was significant across the three tasks (β = 0.34, *SE* = 0.03, t = 10.47, $p < .001$). There was a significant interaction between the Group and Task variables. The group difference in lexical access task was significantly larger than in the grammatical processing task (β = -0.13, *SE* = 0.04, t = -3.19, $p < .01$) and semantic processing task (β = -0.17, *SE* = 0.04, t = -4.04, $p < .001$). As evident from this interaction, the largest processing speed difference between native and nonnative participants lied in the lexical access task.

Processing stability

The processing stability model showed significant task effects (see Appendix Table A3). Native participants' reaction times were more stable than those of the nonnatives in the language processing tests ($\beta = 0.06$, *SE* = 0.01, $t = 6.15$, $p < .001$). The CV model also showed a significant interaction between Group and Task variables. Compared to the lexical access task, the group difference decreased significantly in the grammatical processing task ($\beta = -0.10$, *SE* = 0.02, $t = -5.48$, $p < .001$) and the semantic processing task ($\beta = -0.04$, *SE* = 0.02, $t = -2.14$, $p < .05$). Similar to the RT model, this interaction indicated that the difference between native and nonnative groups in processing stability also peaked in the lexical access task.

Vocabulary size

According to the vocabulary model (see Appendix Table A4), the native group's auditory vocabulary size was larger than that of the nonnative groups ($\beta = -74.61$, *SE* = 5.65, $t = -13.21$, $p < .001$).

2.3.2 Comparisons among nonnative groups

Processing accuracy

The processing accuracy model for the nonnative groups did not show significant task effects (see Appendix Table A5). The AH-regular group had significantly lower accuracy than the SA-onset group across the language processing tasks ($\beta = -0.88$, *SE* = 0.12, $z = -7.10$, $p < .001$). Accuracy of the SA-onset and AH-intensive groups did not differ across tasks ($\beta = -0.15$, *SE* = 0.13, $z = -1.19$, $p > .05$), nor for any particular task ($\beta = 0.22$, *SE* = 0.17, $z = 1.29$, $p > .05$; $\beta = 0.22$, *SE* = 0.21, $z = 1.03$, $p > .05$). The interaction between Group and Task variables in the accuracy model reached significance for the comparison between the AH-regular and SA-onset groups. The advantage of the SA-onset group over the AH-regular group in the lexical access task was significantly larger than in the grammatical processing task ($\beta = 0.45$, *SE* = 0.16, $z = 2.77$, $p < .01$), but did not differ significantly from that in the semantic processing task ($\beta = -0.33$, *SE* = 0.19, $z = -1.70$, $p > .05$). Therefore, compared to the other tasks, the grammatical processing task showed the smallest difference between these two groups.

In summary, the SA-onset and AH-intensive groups had similar accuracy levels across the three tasks, while the AH-regular group had the lowest accuracy among the nonnative groups. Compared to other tasks, the grammatical processing task could least distinguish the three nonnative groups.

Processing speed

The processing speed model for the nonnative groups showed significant task effects (see Appendix Table A6). The AH-regular group responded significantly more slowly than the SA-onset group across the three processing tasks ($\beta = 0.10$, *SE* = 0.03, $t = 4.02$, $p < .001$). However, the differences between SA-onset and AH-intensive groups in reaction time failed to reach significance across the three tasks ($\beta = -0.00$, *SE* = 0.03, $t = -0.10$, $p > .05$), and they did not interact with task effects. The interaction between Group and Task variables was significant for the difference between AH-regular and SA-onset groups: this group difference was significantly larger in the lexical access task than in the grammatical processing task ($\beta = -0.08$, *SE* = 0.03, $t = -2.76$, $p < .01$), but did not differ significantly from that of the semantic processing task ($\beta = 0.03$, *SE* = 0.03, $t = 1.11$, $p > .05$). Therefore, similar to the accuracy model, the RT model showed that the grammatical processing task manifested the least difference between AH-regular and SA-onset groups.

The results for processing speed and accuracy were consistent across the three tasks: performance of the AH-regular group was the least accurate and the slowest, but the AH-intensive and SA-onset groups did not differ in either accuracy or speed in the three processing tasks. The nonnative groups showed the smallest differences in the grammatical processing task.

Processing stability

The processing stability model for the nonnative groups showed significant task effects (see Appendix Table A7). However, the SA-onset and AH-regular groups did not generally differ in processing stability across the processing tasks ($\beta = 0.02$, *SE* = 0.01, $t = 1.68$, $p >.05$). The SA-onset and AH-intensive groups did not differ across the three tasks either ($\beta = 0.01$, *SE* = 0.01, $t = 1.36$, $p > .05$). There was no significant interaction between the Group and Task variables. Therefore, there was no difference among the three nonnative groups in processing stability, or processing stability may not be a sensitive measure to differentiate the nonnative groups in this study.

Vocabulary size

According to the vocabulary size model of the nonnative groups (see Appendix Table A8), the SA-onset group had significantly larger vocabulary size than the AH-regular group (β = -16.19, *SE* = 4.47, *t* = -3.62, *p* < .001) and smaller vocabulary size than the AH-intensive group (β = 15.27, *SE* = 4.65, *t* = 3.28, *p* < .01).

2.4 Discussion

2.4.1 Differences between natives and nonnatives in vocabulary size

Our first research question addressed the difference in vocabulary size between native and nonnative participants. As expected, the results show that the native participants had larger listening vocabulary than the nonnative. This aligns with findings from L2 reading research (e.g., Dąbrowska, 2019). Dąbrowska compared native and nonnative performance using the Vocabulary Size Test (Nation & Beglar, 2007), which measured receptive vocabulary in reading, and found a between-group difference of medium effect size (*r* = 0.35). However, the listening vocabulary size difference between natives and nonnatives in the present study (*r* = 0.7) seems to be larger than that of Dąbrowska's study. One of the reasons for the effect size difference may be that EFL learners' listening vocabulary is usually smaller than their reading vocabulary, as EFL learners acquire vocabulary mostly through reading rather than listening. According to Sadoski's (2006) dual-coding view of vocabulary learning, written and spoken words are encoded separately, forming visual and auditory mental representations, respectively. When learning words solely from reading, L2 learners may vocalize words and create their sound representations in a deviant form. Consequently, upon encountering the correctly pronounced forms, they may not recognize them during listening activities. Therefore, knowledge acquired in reading may not necessarily generalize to listening (see also DeKeyser, 2007; Rodgers, 2011 about skill specificity in L2 learning).

2.4.2 Differences between natives and nonnatives in processing efficiency

Our first research question also concerned whether native speakers outperform advanced L2 learners in language processing efficiency. The results show that native participants' responses in the three speeded judgment tasks were more accurate, faster and more stable than those of the nonnative. This yields further evidence that L2 processing tends to be erroneous, slow and effortful compared to L1 processing (Cop, Drieghe, & Duyck, 2015; Hahne, 2001).

The present results show that the most prominent difference in language processing between native and nonnative groups lies in speed of word recognition. Such an L2 disadvantage in word recognition could be due to multiple factors (Weber & Broersma, 2012). First and foremost, usage-based theories of second language acquisition emphasize the role of token frequency in lexical entrenchment (e.g., Bybee, 2006; Schmid, 2007). The failure of EFL learners to achieve more efficient word recognition might be partly due to the low token frequency of lexical items in their input compared to naturalistic input. Similarly, bilinguals may have weaker links between word forms and semantics than monolinguals due to less experience with their L2 lexical items (see Gollan, Montoya, Cera, & Sandoval, 2008), resulting in slower word recognition (Dijkgraaf, Hartsuiker, & Duyck, 2019). Secondly, for nonnatives, distinguishing L2 phoneme contrasts that do not exist in the learner's L1 (e.g., /iː/ vs. /ɪ/, /l/ vs. /r/, and /eɪ/ vs. /e/) is difficult. Such phoneme contrasts were also present in our testing stimuli (e.g., sheep vs. ship; doll vs. door; whale vs. well). Consequently, strong activation of neighbouring phonemes will complicate lexical selection for L2 listeners (Weber & Cutler, 2004). Thirdly, the set of competing word candidates in L2 listening may be larger than in L1 listening due to cross-language interference. L2 listeners will thus face more difficulty deactivating unintended word candidates than natives (Weber & Broersma, 2012). However, it should be noted that, since the three language processing tasks were administered in a fixed order, task difficulty may be confounded by administration order, especially for the nonnatives.

2.4.3 Effects of EFL exposure on vocabulary size in EFL learners

The second research question addressed whether and how EFL learners from three different learning contexts differed in vocabulary size. The AH-intensive group was found to have significantly larger vocabulary size than the SA-onset group, who in turn outperformed the AH-regular group in vocabulary size. These findings suggest that vocabulary size increases significantly along with EFL exposure, which agrees with usage-based theories (e.g., N. Ellis, 2006) on the role of exposure in vocabulary acquisition. Demonstrating the facilitatory effect of exposure on acquiring listening vocabulary, this result complements L2 vocabulary acquisition research on the correlation between exposure and reading vocabulary size (e.g., Milton, 2009). This result seems to support the view that vocabulary acquisition relies on general learning mechanisms and, given enough frequency of target language exposure, more L2 vocabulary can be acquired over time and practice.

2.4.4 Effects of EFL exposure on processing efficiency in EFL learners

The second research question also addressed whether EFL learners from three different learning contexts differed in processing efficiency. The SA-onset group gave faster and more accurate responses than the AH-regular group across the three speeded tasks, but did not differ from the AH-intensive group in any of the measures in these tasks. The results are only partly in line with our initial expectation on the relationship between L2 exposure and processing efficiency. On the one hand, as expected, the significant difference between AH-regular and SA-onset groups indicates that more language exposure is associated with faster and more accurate lexical access, greater grammar sensitivity and better ability to form semantic propositions in listening (for similar findings on first language acquisition, see Chateau & Jared, 2000; Dijkgraaf, Hartsuiker, & Duyck, 2019; Wells, Christiansen, Race, Acheson, & MacDonald, 2009). On the other hand, the lack of a significant difference between SA-onset and AH-intensive groups is somewhat unexpected. This result suggests that more language exposure in an EFL learning context does not necessarily lead to higher language processing efficiency. Note that the performance of nonnative groups in this study was still quite distant from ceiling performance as suggested by native performance, such that there still seemed to be ample room for improvement. Building on usage-based theories (e.g., Bybee, 2003, 2006; N. Ellis, 2006; Schmid, 2007), we expected that the nonnative participants with more exposure to the target language should have higher processing efficiency than those exposed less. This was, however, not supported by the comparison between the SA-onset and AH-intensive groups.

In order to explain the lack of a significant difference between the SA-onset and AH-intensive groups in processing efficiency, we need to take into account not only amount of exposure but also type and quality of exposure. One possible explanation for our results pattern is that, similar to the well-established fossilization phenomenon in second language acquisition (e.g., Han, 2013; Han & Odlin, 2006; Selinker, 1972), language exposure in EFL contexts may have a limited effect on improving the learner's processing abilities beyond a certain level of proficiency. EFL learning contexts may often only provide impoverished and limited L2 input, in the form of textbooks and similar instructional materials. Such instructional materials tend to cover a large number of word types without guaranteeing sufficient repetition of word tokens (Matsuoka & Hirsh, 2010; Sun & Dang, 2020), which is not beneficial for L2 lexical entrenchment. Note that more deeply entrenched knowledge is assumed to be processed more automatically. Moreover, EFL learning contexts may be characterized by an overreliance on visually presented input, which presents a stark contrast to the situation of naturalistic first-language acquisition where auditory input is dominant. Due to its transient nature, EFL learners often find speech more difficult to process than text in the target language. It is not rare to encounter foreign-language learners who can read well but struggle while listening. These aspects of EFL exposure may hamper the development of L2 processing efficiency in listening comprehension. What is implied is that learners, especially those at intermediate and advanced proficiency levels, may find it harder (or even are unable) to utilize language exposure in EFL contexts for the purpose of improving their proficiency, especially processing efficiency. They may have to change their language learning methods or environments for further progress.

The present study also shows that grammar sensitivity, out of the processing components under investigation, is the area that least distinguishes the three nonnative groups. According to usage-based theories, type frequency determines the productivity of linguisitc constructions (Bybee & Hopper, 2001; Bybee & Thompson, 2000; N. Ellis, 2002, 2006). The lack of grammar sensitivity may be attributed to the relatively low type frequency of grammatical constructions in EFL exposure. This result may also be explained by Clahsen and Felser's (2006) Shallow Structure Hypothesis, which argues that "the syntactic representations adult L2 learners compute during comprehension are shallower and less detailed than those of native speakers". Comparing the performance of mature native speakers and L2 learners in the domains of morphology and syntax, their study found that adult L2 learners tend to rely on lexical-semantic information and overlook syntactic cues during sentence processing. This characteristic of nonnative processing determines that it is very difficult for L2 learners to develop grammar sensitivity. It may, therefore, not be surprising if increasing exposure does not lead to much progress in grammatical processing. As mentioned previously, we should bear in mind that task administration order may potentially confound participants' task performance when interpreting this result.

Last but not least, the three nonnative groups differed in processing speed, but not in their processing stability. Based on Segalowitz's (2010) distinction between quantitative and qualitative changes in learners' language processing mechanisms, we take these results to indicate that the proficiency differences between groups mainly involve a general speeding-up of controlled performance without a restructuring of the underlying processing mechanisms.

2.4.5 Pedagogical implications

Firstly, if we compare our results to those of Dąbrowska (2019), it seems that the difference between natives and nonnatives in listening vocabulary is larger than that of reading vocabulary, which may suggest an imbalance in how reading and listening develop among EFL learners. This imbalance may be partly due to the transient nature of auditorily presented materials. Learners and teachers should be encouraged to adopt new methods and techniques in the language classroom (e.g., electronic textbooks combining text and audio) that may help make auditory input more accessible and analyzable.

Secondly, the failure of EFL learners to achieve more efficient word recognition and grammatical processing might be partly due to the characteristics of EFL input distribution, namely low token frequency of lexical items and low type frequency of grammatical constructions. Thus, efforts should be put into creating more opportunities (e.g., online learning games or naturalistic conversations) for EFL learners to deeply entrench their word representations and subsequent word recognition. Teachers may need to explore ways of optimizing L2 input (e.g., using focus-on-form techniques; see R. Ellis, 2016) to help improve learners' grammar sensitivity.

Thirdly, increasing amount of EFL exposure leads to larger vocabulary size but not necessarily to higher processing efficiency, even in a stage where there is still ample room for improvement. The limited effect of EFL exposure on improving processing efficiency may be partly attributed to their instructed language education, which usually emphasizes literacy over oral communication (Ruan & Leung, 2012). Consequently, EFL learners tend to emphasize knowledge accumulation at the expense of processing automatization. A shift of educational focus from literacy to oral communication might help alleviate this problem to some extent.

Finally, accurate measurement of processing efficiency indexes learners' L2 strengths and weaknesses in language use, thus providing learners, teachers, and researchers more insight in SLA. Therefore, testing processing efficiency might be a useful addition to current L2 learning and teaching practice.

2.4.6 Limitations and future recommendations

One limitation of this study is that there is no objective and fine-grained measurement of participants' actual language exposure in each group due to the difficulty of quantifying L2 input. Consequently, the exact difference in language exposure between the nonnative groups was unknown. However, judging from the estimation of instructed and self-directed exposure as well as the result patterns of our vocabulary test, we do think our statement that exposure levels of the SA-onset group were in between those of the other two groups is justified. Another limitation of our study is that our tasks were administered in a fixed order, such that between-task differences in participants' performance may be confounded with potential task-order effects. This may weaken some of our findings drawn on between-tasks comparisons, as discussed previously. We recommend future research to counterbalance over tasks if interested in examining whether certain target effects are task-specific or not.

2.5 Conclusions

In sum, this study examines L2 listening proficiency in EFL learners from different learning contexts, associated with different exposure levels to English. Regarding auditory vocabulary knowledge, the at-home-intensive group outperformed the study-abroad-onset group, who in turn outperformed the at-home-regular group. Regarding spoken-language processing efficiency, the at-home-intensive and study-abroad-onset groups did not differ in any of the processing measures, but they both outperformed the at-home-regular group in accuracy and speed of processing across the processing tasks. The comparison between the AH-regular and SA-onset groups shows that the non-English majors achieved larger vocabulary and higher processing efficiency with more exposure to the target language, which may relate to their preparation for going abroad. However, the comparison between SA-onset and AH-intensive groups shows that the higher exposure that English majors received only results in larger vocabulary size, relative to the international non-English majors, but did not result in higher language processing efficiency. That is, increasing amount of EFL exposure was associated with larger vocabulary size but not always with higher processing efficiency. This study suggests that language exposure in EFL contexts has a limited effect on developing language processing efficiency. Knowledge accumulation and processing automatization may not be equally affected by EFL exposure. Lastly, note that this study mostly involves EFL learning contexts in China, while L2 learning contexts vary considerably worldwide. Future research should incorporate a cross-context comparison between EFL learning contexts and naturalistic learning contexts, especially studying abroad contexts, to gain a more comprehensive understanding of second language acquisition.

Chapter 3: The effect of learning context on L2 listening development: Knowledge and processing

Abstract

Little research has been done on the effect of learning context on L2 listening development. Motivated by DeKeyser's (2015) skill acquisition theory of second language acquisition, this study compares L2 listening development in study abroad (SA) and at home (AH) contexts from both language knowledge and processing perspectives. 149 Chinese postgraduates studying in either China or the UK participated in a battery of listening tasks at the beginning and at the end of an academic year. These tasks measure auditory vocabulary knowledge and listening processing efficiency (i.e., accuracy, speed, and stability of processing) in word recognition, grammatical processing, and semantic analysis. Results show that, provided equal starting levels, the SA learners made more progress than the AH learners in speed of processing across the language processing tasks, with less clear results for vocabulary acquisition. Studying abroad may be an effective intervention for L2 learning, especially in terms of processing speed.

This chapter has been adapted from

Yu, X., Janse, E., & Schoonen, R. (2020). The effect of learning contexts on L2 listening development: Knowledge and processing. *Studies in Second Language Acquisition,* 1–26. doi:10.1017/S0272263120000534. (First View)

3.1 Introduction

Second language (L2) learning contexts vary widely in quality and quantity of input, output, and interaction, inevitably leading to different L2 development patterns and attainments. Cross-context comparisons, e.g., comparing study-abroad (SA) contexts and at-home (AH) contexts, may help to reveal the unique characteristics of L2 development in different learning contexts (Kroll, Dussias, & Bajo, 2018). L2 development can be examined from two distinct perspectives, i.e., language knowledge (e.g., vocabulary and grammar) and processing skills (e.g., how rapidly or easily one can understand a sentence). As argued by DeKeyser (2007, 2015) and Hulstijn, Van Gelderen, and Schoonen (2009), L2 learners have to accumulate knowledge of the target language, as well as improve the efficiency with which that knowledge can be processed. Previous studies (e.g., Collentine, 2004; Freed, Segalowitz, & Dewey, 2004; Håkansson & Norrby, 2010; Pliatsikas & Marinis, 2013a, 2013b; Sasaki, 2007; Segalowitz & Freed, 2004) have investigated how learning context affects the acquisition of both knowledge and processing aspects of language proficiency. For instance, Collentine (2004) analysed speech produced by American learners of Spanish (in study-abroad and regular-classroom settings) in an Oral Proficiency Interview before and after a semester. All participants were university students with no previous contact with Spanish. The results showed that formal education in an AH context facilitated development of discrete grammatical and lexical knowledge (indicated by use of grammatically marked forms and lexical frequencies respectively), while the immersion in an SA context was beneficial for the development of oral fluency, a form of processing ability in speaking.

However, studies of L2 learning contexts have rarely focused on listening comprehension, an important but often neglected area of second language acquisition. The few listening studies that have investigated L2 learning contexts (e.g., Cubillos, Chieffo, & Fan, 2008; Llanes & Muñoz, 2009) have largely employed holistic measurement methods (e.g., spoken passage comprehension) to test listening proficiency. Consequently, our knowledge of the effect of learning contexts on the finer components of listening comprehension is rather limited. Therefore, it is not known whether study-abroad learners will be able to recognize more words, or whether they will be faster in recognizing a word, processing grammatical information and forming the semantic proposition of a sentence in listening comprehension than their stay-at-home peers. To fill this gap, this study takes a componential view of listening proficiency, in terms of auditory vocabulary knowledge and listening processing efficiency, and examines L2 listening development in SA and AH learning contexts for an academic year. By comparing study-abroad learners with their stay-at-home peers, we aim to investigate whether and to what extent a shift of learning context from AH to SA is an effective intervention for improving adult L2 learners' listening proficiency.

3.1.1 Study-abroad vs. at-home learning contexts

Studying abroad is often considered as the best L2 learning context (Freed, 1995), as it usually involves a language environment shift where learners have to inhibit the first language and immerse themselves in the target language (Jacobs, Fricke, & Kroll, 2016; Linck, Kroll, & Sunderman, 2009). Studying in a country where students' L2 is spoken as the native language guarantees abundant native input, opportunities for output, informative feedback, and interaction. Contrastively, even though nowadays it is relatively easy to have access to authentic L2 input through digital platforms (e.g., Netflix, Audible, and Apple News), AH contexts may be criticized for relatively inadequate L2 exposure, over-reliance on rote learning, and limited opportunities for interactive communication. The potential advantages of SA contexts over AH contexts argue for the linguistic benefits of studying abroad. In an ideal scenario where learners are fully immersed in their new environment, SA contexts seem to solve the problems leading to generally low attainments of adult L2 learners in AH contexts. However, SA experiences are hardly ever ideal and immersion degrees may sometimes be overestimated. Newcomers usually experience a gradual process of socialization, starting with compatriots, then expanding to other international students, and finally to locals (Coleman, 2013). As communication across linguistic and cultural boundaries is challenging, the socialization process may stagnate at any stage, as international students may tend to foreground their national identity against intercultural identities during the intercultural experience (Maeder-Qian, 2018). It is not rare that same-nationality students clutter in and out of class. Consequently, the linguistic impact of studying abroad may be compromised due to integration problems.

During the past two decades, blooming international education has led to multiple study-abroad studies (e.g., Dwyer, 2004; Leong, 2007; Sasaki, 2011; Segalowitz & Freed, 2004; Williams, 2005). These studies have been set up from various research angles, focusing on issues such as language proficiency, cross-cultural competencies, personality changes, and career growth. Varela (2017) performed a meta-analysis of 33 studies on language development in study-abroad learners, focusing on dependent variables such as general language proficiency, written proficiency, vocabulary size, and speech rate. Varela reported a large effect of studying abroad on enhancing L2 proficiency ($d = 0.975$). However, this meta-analysis was largely limited to comparing pre- and post-test differences in the SA context. By comparing with Plonsky's (2011) meta-analysis on language learning in the AH context ($d = 0.55$), Varela claimed that study-abroad programmes facilitated second language acquisition. Note that this comparison is not direct. As for individual studies that directly contrasted SA and AH contexts, some found that SA learners had greater gains in knowledge of nativelike language usage (Foster, Bolibaugh, & Kotula, 2014), use of communication strategies (Lafford, 2004), grammar (Marqués-Pascual, 2011; Pliatsikas & Marinis, 2013b), accent (Martinsen, Alvord, & Tanner, 2014), pragmatic competence (Matsumura, 2001), writing proficiency (Sasaki, 2011), and oral proficiency (Segalowitz & Freed, 2004), but others reported marginal or no differences as a function of learning context in terms of grammar (Isabelli-García, 2010;

Pliatsikas & Marinis, 2013a), and pragmatic comprehension (Taguchi, 2010). Mixed outcomes may relate to the fact that different studies focused on different aspects of language acquisition, which may not be equally sensitive to the effect of learning context.

Furthermore, previous studies have investigated the effects of learning context on fluency, accuracy, and complexity of L2 oral and written production. Learners' oral fluency, measured by speech rate and mean run length with no pause, usually improves after studying abroad (Mora & Valls-Ferrer, 2012; Segalowitz & Freed, 2004). However, accuracy and complexity measures of oral production, such as frequency of errors, length and syntactic complexity of sentences, shows conflicting results across studies. Some studies provided evidence that study-abroad groups show gains in L2 grammatical complexity and accuracy relative to AH groups (Håkansson & Norrby, 2010; Howard, 2001; Llanes & Muñoz, 2013; Marqués-Pascual, 2011). Other studies, however, claimed that learners achieved better fluency only by using appropriate fillers, modifiers, formulae, and compensation strategies, while their grammatical competence remained unchanged (see Collentine, 2004; DeKeyser, 1991). Similarly, for written production, the benefits of studying abroad have usually been reported to manifest on writing fluency but not necessarily on measures of accuracy and complexity (see Knoch, Rouhshda, Oon, & Storch, 2015; Sasaki, 2007). These results together seem to suggest that learning contexts have differential effects on different aspects of L2 production. Yet it remains unclear how and to what extent studying abroad affects various aspects of L2 listening comprehension. We set out to evaluate the impact of an SA context, in comparison to AH contexts, on L2 listening development in terms of auditory vocabulary knowledge and processing efficiency.

3.1.2 Vocabulary knowledge

As a type of declarative knowledge, L2 vocabulary knowledge can be acquired either incidentally (i.e., through reading and listening activities aimed at communication and not explicitly at vocabulary learning), or intentionally (i.e., through deliberate memorization of lexical information in order to enlarge vocabulary size of a target language). *Incidental learning* is widely held to be the major source for accumulating vocabulary knowledge in both L1 and L2 learners, whereas only a relatively small amount of vocabulary is acquired via *intentional learning* (Hulstijn, 2003). Studies of incidental learning reported significant vocabulary gains through extensive reading by L2 learners (Swanborn & De Glopper, 2002; Horst, 2005; Pellicer-Sánchez & Schmitt, 2010). However, L2 incidental vocabulary learning has been associated with low retention rates, which is why some studies (e.g., Horst, Cobb, & Meara, 1998; Waring & Takaki, 2003) have claimed that the role incidental learning plays in L2 vocabulary acquisition may have been overestimated. Intentional learning, on the other hand, has been found to be much more effective than incidental learning in retaining lexical information, especially over a short period of time (e.g., Schmitt, 2008; Swanborn & De Glopper, 1999).

SA contexts may be superior to AH contexts in facilitating vocabulary acquisition for several reasons. Firstly, the naturalistic exposure in SA contexts arguably guarantees more opportunities for incidental vocabulary learning than AH contexts, with the latter likely being explicitly geared to intentional learning by memorization. Second, interaction and negotiation of meaning have been found to facilitate L2 vocabulary acquisition (Ellis, Tanaka, Yamazaki, 1994; Long, 1996; Newton, 2013). Through negotiating meaning (e.g., by rephrasing or asking for clarification), learners and their interlocutors overcome comprehension difficulties, which may then become learning opportunities. Immersion in SA contexts allows learners to interact and negotiate in the target language, while regular classroom settings in AH contexts are often criticized for limited opportunities for interaction. Thirdly, according to Mayer's (2009) Cognitive Theory of Multimedia Learning, the brain integrates information from visual and auditory channels (e.g., words, pictures and auditory information) to create mental representations. If visual and auditory channels provide congruent information, people learn better from both channels than only from one channel. Learning words solely from reading, therefore, may not be as efficient as in combination with their auditory form. L2 learners may vocalize words acquired via reading in a deviant form, and consequently may not recognize them in the correctly-pronounced form in listening activities. Learners from AH contexts often suffer from the lack of auditory form because listening proficiency is not capitalised on in educational settings, and this problem is supposed to be reduced to some degree for SA learners immersed in their L2. The question as to whether SA contexts better facilitate vocabulary acquisition than AH contexts boils down to whether and to what extent L2 learners can benefit from the incidental learning opportunities provided by an SA context.

Previous studies have compared vocabulary acquisition across different learning contexts (DeKeyser, 1991; Dewey, 2008; Ife, Vives Boix, & Meara, 2000; Llanes & Muñoz, 2009; Milton & Meara,1995). Milton and Meara (1995), for instance, compared students' half-yearly vocabulary growth before and after the onset of their six-month study-abroad program, and found that the average growth rate in an SA context was four times bigger than that in an AH context. However, Dewey (2008) compared vocabulary gains made by intermediate English learners of Japanese in three learning contexts during nine-to-thirteen weeks with various vocabulary tests. Participants were either in a study-abroad program, in an intensive domestic immersion program, or in a formal classroom setting. Dewey (2008) found that vocabulary gains in the SA context were not significantly different from those in the intensive domestic immersion setting. This study suggested that the benefits of SA contexts on vocabulary acquisition might stem only from the *amount* of language exposure, rather than the difference in learning contexts. Note that most of the above-mentioned studies tested vocabulary knowledge in reading. Our knowledge about the relationship between learning contexts and auditory vocabulary, which is related to but different from reading vocabulary, is rather limited. Therefore, this study compares auditory vocabulary acquisition in SA, AH regular classroom and AH intensive instruction settings to examine the relationship between auditory vocabulary acquisition and learning context.

3.1.3 Processing efficiency

According to DeKeyser's (2015) skill acquisition theory of second language acquisition, learners go through declarative, procedural and automatic stages sequentially during the acquisition process. Firstly, L2 learners start with explicitly learning declarative knowledge (e.g., vocabulary and grammar). Secondly, learners develop procedural knowledge, which is the knowledge exercised in the accomplishment of a task, after a few practice trials. This proceduralization of knowledge is realized when the execution of a target performance gets routinized or chunked (Anderson, 2007; Taatgen & Lee, 2003). This procedural stage is not particularly time-consuming (DeKeyser, 1997). Finally, in order to use language spontaneously or effortlessly, learners need a large amount of practice to automatize the procedural knowledge acquired in the previous stage. During this slow and gradual process of automatization, learners will comprehend and produce language in a more rapid way, showing fewer errors and requiring less attention. As the automatization of L2 processing is slow (DeKeyser, 2015; Lim & Godfroid, 2015), it is difficult to observe rapid progress in regular foreign language classroom settings. However, the dramatic environmental shift entailed by a study abroad experience may accelerate L2 automatization, thus creating a situation to test hypotheses about the development of L2 processing skills over a relatively short period of time. This study describes L2 processing skills during listening comprehension in terms of processing efficiency, to avoid the (related, but theoretically charged) term "automaticity" (for a review on automaticity, see Segalowitz, 2003, 2010). Processing efficiency is operationalised as a multi-dimensional construct comprising accuracy, speed, and stability of processing.

Multiple previous studies of second language processing have compared L1 and L2 processing in word recognition, parsing, semantic or phonological processing (for a review, see Jiang, 2018). Rather than contrasting L1 and L2 processing, only a few studies investigated the development of L2 processing in relation to language learning contexts. For example, Segalowitz and Freed (2004) compared the performance of English (SA and AH) learners of Spanish in a semantic classification task before and after one semester. They reported no effect of learning context on lexical access (quantified as speed of semantic classification decisions). As for studies on grammar acquisition, Isabelli-García (2010) investigated gender acquisition (using grammaticality judgment tests) in intermediate English (SA and AH) learners of Spanish over four months, and Pliatsikas and Marinis (2013a) studied the processing of past tense morphology in highly proficient Greek learners of English (one group with over-a-year SA experience and another with only regular AH classroom exposure) with a self-paced reading task. Both studies reported no effect of learning context on grammar acquisition (i.e., gender agreement and past tense, respectively). However, these studies all measure L2 processing capacities in reading, an activity that is emphasized and well-practiced in AH learning contexts. It is not clear how the various cognitive processing abilities in L2 listening comprehension develop across learning contexts. This study investigates the effect of learning contexts on processing abilities at three different levels of listening comprehension, i.e., lexical, morphosyntactic, and semantic levels. More specifically, a series of tasks have been devised to measure L2 processing efficiency in lexical access (e.g., recognizing a spoken word), grammatical processing (e.g., capturing a grammatical feature of an utterance), and semantic processing (e.g., understanding the semantic meaning of an utterance). These three processes are critical building blocks towards successful language comprehension (for language comprehension models, see Anderson, 2015; Cutler and Clifton, 1999; Goss, 1982).

3.1.4 Current study

We hypothesize that advanced English learners who study abroad and have experienced a shift of language environment from an English as a Foreign Language (EFL) country (China) to an English as a Native Language (ENL) country (UK) will make more progress than their domestic counterparts in terms of both vocabulary size and language processing efficiency. To test this hypothesis, we invited Chinese international non-English-major postgraduates studying in the UK (SA group, with no previous study-abroad experience), Chinese domestic English-major postgraduates (AH-intensive group) and domestic non-English-major postgraduates (AH-regular group) to participate in a series of English tests at the beginning of their postgraduate program and again after one academic year.

With respect to baseline proficiency, the SA group can be expected to be similar to the AH-regular group since both groups were majoring in non-English subjects. At the same time, the SA group had done intensive preparation, including intensive English learning, in order to qualify for studying abroad. Consequently, their baseline proficiency may also turn out to be more similar to that of the AH-intensive group. As baseline language proficiency has been shown to relate to size of language learning gains over time (Davidson, 2010; Brecht & Robinson, 1995), both AH groups were included as reference groups for the SA group, to provide a more complete comparison between study-abroad and at-home learning contexts.

The pre-test and post-test design allows us to compare L2 listening proficiency improvement, in terms of processing efficiency and auditory vocabulary size, across different learning contexts. To investigate the effect of learning contexts on L2 listening development, this study sets out to answer the following questions:

1) Does the SA group show more improvement in auditory vocabulary size than the two AH groups over the course of an academic year?

2) Does the SA group show more improvement in language processing efficiency (i.e., accuracy, speed, and stability of processing) than the two AH groups over the course of an academic year? And if so, is the group difference in improvement constrained to specific linguistic abilities (i.e., lexical access, grammatical processing, and semantic processing)?

3.2 Methodology

3.2.1 Participants

149 Chinese postgraduates studying abroad or domestically took the pre-test and post-test with an interval of seven months. Among them, there were 47 non-English majors studying in the UK (SA group), 53 non-English majors studying in China (AH-regular group), and 49 English majors studying in China (AH-intensive group). All the participants finished bachelor education in China, with no previous study-abroad experience before the pre-test. The SA group is the target group representing a rapidly increasing population of L2 learners who start learning English as an FL at home and later on move to an English-speaking country as an adult to participate in an SA program. The two AH groups are both control groups to be compared with the SA group in order to compare English proficiency improvement across different learning contexts.

Table 3.1: Background information of these participant groups: AH-regular group (*N* = 53), AH-intensive group (*N* = 49), and SA group (*N* = 47).

Group	Mean age (*SD*)[1]	Major	Hours of English instruction during Bachelor program (before pre-test)[2]	Hour of English instruction during Master program (between pre- and post-test)[2]	Mean standardized test score (*SD*)[3]
AH-regular	23.2 (1.0)	Non-English	144 hours	72 hours	542 (32.6)
AH-intensive	23.4 (1.4)	English	1620 hours	468 hours	563 (38.8)

SA	23.3 (1.1)	Non-English	144 hours	420 hours	528 (54.7)

Notes: 1. All participants reported their age in the post-test questionnaire, while this information was not complete in the pre-test questionnaire. The means and SDs of Age hereof were therefore calculated based on the post-test questionnaire for the sake of data completeness.

2. Hours of instruction were calculated based on the credits required by each educational program.

3. Not all nonnative participants reported their CET-6 (i.e., a national standardized English test) score. The means and SDs for this variable are therefore based on smaller sample sizes (i.e., 147 out of 165). An ANOVA test shows that the AH-regular and the SA group did not differ significantly (p = .221), while the AH-intensive group significantly outperformed the AH-regular (p = .03) and the SA (p = .0003) group in CET-6. However, these results have to be interpreted with caution because participants took the CET-6 test up to three years before the pre-test, and hence maybe well before the SA group started to prepare for studying abroad.

Before our pre-test, the English courses of the AH-intensive group included Comprehensive English, Oral English, Listening Comprehension, Intensive Reading, Extensive Reading, Writing, Literature, Linguistics (around 1620 hours in total prescribed by their bachelor programs). The AH-regular group and SA group only had a College English class once per week (around 144 hours prescribed by their bachelor program). According to their standardized test scores (see Table 3.1), at some point during their bachelor program, the AH-regular and the SA groups did not differ in L2 proficiency, while the AH-intensive group had significantly higher language proficiency than the AH-regular and the SA group. Afterwards, the SA group, nevertheless, was expected to have learned English (mainly out of class) more than the AH-regular group since they had to prepare for studying abroad. Therefore, we estimated that the baseline language proficiency of the SA group *at pre-test* may be somewhere between that of the AH-regular (non-English-major) and AH-intensive (English-major) group.

Between our pre- and post-test, the AH-intensive group had English-medium courses (given by Chinese teachers), which included Literature, Linguistics, Translation, Interpretation, Methodology (around 468 hours in total), but no basic language learning courses. The SA group also had English-medium courses (around 420 hours in total) but no basic language learning course. The AH-regular group had a two-hour college English class every week (around 72 hours in total). Therefore, though the SA group was not majoring in English, both the SA and the AH-intensive group would have English-medium education in the coming academic year. Contrastively, only the AH-regular group would mainly have Chinese-medium education.

3.2.2 Materials

The testing materials include a lexical access task, a grammatical processing task, a semantic processing task, and the Peabody Picture Vocabulary Test Fourth Edition (PPVT™-4).

The first three tasks were timed decision tasks used to measure language processing efficiency at the lexical, morphosyntactic, and semantic level respectively. Measures of these tasks were accuracy, reaction time (RT), and coefficient of variation (CV), indicating accuracy, speed, and stability of processing respectively. These tasks are designed to focus on how efficiently learners can process their L2 (e.g., how rapidly they can recognize a word, how easily they process a certain grammatical structure, or how fast they can understand an utterance), in addition to their accuracy in performing these tasks. We specifically aimed to minimize the effects of limited vocabulary or grammar knowledge when performing language processing tasks. L2 listening tests often involve holistic measurement methods like spoken passage comprehension. Scores thereof are used as a general indicator of listening proficiency, but reflect little about listening effort or about different components of the listening process. Test stimuli in our tasks are carefully manipulated to allow for detecting the subtle nuances in language processing that are not easily identifiable by existing standardized tests. Therefore, the use of these speeded-response tasks will allow us to investigate the effect of learning context on the development of L2 processing skills.

The (untimed) PPVT™-4 was used to measure auditory vocabulary size, a form of declarative knowledge. The measure of this task was a score for the number of vocabulary items correctly identified, functioning as an approximation of auditory vocabulary size. The PPVT is designed to measure receptive (auditory) vocabulary size of English native speakers aged from 2:6 to over 90 years. The test's vocabulary items are not restricted to the vocabulary of any specific purpose or discipline. This test is not subject to ceiling effects, either. This format of the test (i.e., choosing a picture that matched the word participants had heard) is more intuitive and straightforward than the commonly-used multiple-choice format that involves choosing among synonyms or descriptions.

Lexical access task

The lexical access task measured how well participants could recognize words with simultaneous presentation of auditory and visual stimuli. In each trial, participants saw a (line-drawing) picture and heard a word simultaneously. They were asked to judge as fast as possible whether the picture and word matched or not by pressing correspondent buttons on a button box. If a picture and a word did not match, their referents shared phonetic similarities (e.g., "kite" and "cat"), fell into the same semantic field (e.g., "apple" and "orange"), or were unrelated (e.g., "frog" and "doctor"). Note that a trial would end and the next trial would start automatically if no response was given within 4 seconds.

This test contained six training trials and sixty experimental trials. Cronbach's alpha indices for the accuracy measures were .6 (pre-test) and .58 (post-test), while those for the RTs were .97 (pre-test) and .96 (post-test). The lower alphas for accuracy are due to the (intended) relative ease of the tasks so that enough valid RTs could be collected.

Grammatical processing task

The grammatical processing task, including six training trials and sixty experimental trials, tapped into the processing of particular grammatical properties. Participants listened to a sentence and saw two pictures simultaneously, and were asked to quickly choose (within eight seconds) the picture that matched the sentence they had heard by pressing a corresponding button on a button box. To make the correct choice, participants had to capture a grammatical cue of that sentence. These grammatical cues could be put into two categories: morpho-syntactic cues, and function words. Morpho-syntactic cues included plural "-s", 3rd person singular "-s", tense, aspect, dative, passive and cleft constructions, and relative clause. For example, participants heard the sentence "the sheep eats" and saw two pictures. In the first picture, there were three sheep eating, while in the second one there was only one sheep eating. The morpheme "-s" in the sentence was the key information leading to the correct picture. For another example, participants heard the sentence "It is the dog that the pig follows" and saw two pictures (one with a dog following a pig, while in the other one there was a pig following a dog). The cleft sentence structure was the syntactic cue in this case. As for function words, this category of sentences contained word-level cues, e.g., prepositions and conjunctions. For example, the participants heard the sentence "The children are marching along the sidewalk" and saw two pictures: in the first one the children were marching across the sidewalk, and in the second one they were marching along the sidewalk. The preposition "along" in the sentence was the function-word cue indicating the correct picture. The function-word category (twenty items out of the total of sixty items) was later excluded for analysis so that this task could better qualify as an indicator of grammatical/morpho-syntactic processing.

The sentence and picture stimuli of this task were partly taken from Kersten (2010), Waters, Caplan, and Rochon (1995) and Weist (2002). We adapted these stimuli into a time-constrained sentence-picture matching test format. Cronbach's alpha indices for the accuracy measures were .57 (pre-test) and .55 (post-test), while those for the RTs were .88 (pre-test) and .89 (post-test).

Semantic processing task

The semantic processing task measured how efficiently participants could form a semantic interpretation of a sentence. Participants were asked to quickly indicate whether a sentence they were listening to was plausible or not. If a sentence was implausible, it violated either obvious factual knowledge (e.g. "A horse is an animal that can fly") or logic (e.g. "If you eat too much, you can get too thin"). The maximum reaction time for each stimulus was set to eight seconds from audio onset. This task, containing six practice trials and fifty experiment trials, originated from Lim and Godfroid (2015). Cronbach's alpha indices for the accuracy measures were .82 (pre-test) and .79 (post-test), while those for the RTs were .95 (pre-test) and .95 (post-test).

Vocabulary size test

The PPVT™-4 (Dunn & Dunn, 2007) was used to measure auditory vocabulary size. Participants heard a word and saw four pictures on a computer screen, and were asked to choose a picture that matched the word they had just heard. A total of 228 test items are grouped into 19 sets of 12 words, which are arranged in order of increasing difficulty. Test administration (20 minutes on average) ended automatically if participants had made more than eight errors in one set. Unlike the first three tasks, the vocabulary size test is not timed. Participants could listen to a word multiple times if necessary. Due to the adaptiveness of the test, participants were administered different subsets of items, which prevented us from computing Cronbach's alpha. The PPVT has a reported reliability of .97 (Dunn & Dunn, 2007).

3.2.3 Procedures

Data collection

The location of data collection was Southeast University at Nanjing, China (for the two AH groups) and Birkbeck, University of London and University College London in the UK (for the SA group). Background questionnaires were sent out beforehand to screen participants for their eligibility. Eligible participants were invited to take the pre-test and the post-test at the beginning and end of an academic year, respectively. After each round of data collection, participants received a small financial reward for their participation.

Data cleaning

Firstly, six items from the grammatical processing and three items from the semantic processing task were excluded as ambiguous items. Items were excluded if they elicited accuracy rates below 80% in native-speaker participants of the previous chapter (Chapter 2). Secondly, two participants from the AH-intensive group were excluded due to either a lower than 50% accuracy rate on any speeded-response test or because their vocabulary score was not within three standard deviations of the group mean. Thirdly, only RTs of valid (i.e., correct) responses were analysed. In addition, RTs below 250 milliseconds (measured from audio onset) were removed as invalid responses.

Statistical analysis

Data analysis was conducted in R version 3.5.1 (R Core Team, 2018). The *glmer* and *lmer* functions in the package *lme4* (Bates, Maechler, Bolker, & Walker, 2015) were used to fit logistic and linear mixed-effects regression models (LMMs), with the optimizer bobyqa. P values were calculated and added into linear regression model outputs with the package *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2017).

Mixed-effects regression models were fitted to predict accuracy (in logit), speed, and stability of responses in the language processing tasks. RTs were log-transformed before model fitting to normalize their distribution. All models took Time, Group and Task as fixed-effects predictors, and included maximal by-time, by-participant, and by-item random intercepts and slopes whenever applicable on the premise of model convergence. As the audio duration of test stimuli and trial number may affect participants' reaction times, these two factors were also entered in the RT model as fixed-effect control variables. Since CVs were calculated as an aggregated measure on task level (i.e., CV = $SD_{RT}/Mean_{RT}$), the CV model did not have any item-level random-effect variable. Finally, a linear mixed-effects regression model, with Time and Group as fixed-effect factors and by-participant random effects, was fitted to predict vocabulary size. For all the models, we sum-coded Group and Task variables to compare the main effects of these variables. However, the Time variable was dummy-coded, with pre-test performance being mapped on the reference level, in order to examine group and task effects at pre-test. Note that the use of mixed-effects regression models allows us to compare between-group differences in terms of their progress over the academic year while statistically controlling for pre-existing between-group differences at pre-test.

3.3 Results

Table 3.2 displays the descriptive statistics of the task performances of the participant groups in the pre- and post-tests. The rest of this section gives a detailed description of the statistical results of the accuracy, RT, CV, and vocabulary size models. For the description of the model results, Group and Task effects at pre-test will be presented before effects concerning the Time variable (i.e., the difference between the pre-test and the post-test), such as the main effect of Time, two-way interactions between Time and Group, and three-way interactions between Time, Group and Task. The interactions between Time and Group, and the interactions between Time, Group, and Task are most critical for answering the research questions of this study.

Table 3.2: Descriptive statistics of task performance in the pre- and post- tests for at-home non-English majors (AH-regular, N = 53), at-home English majors (AH-intensive, N = 49) and for the study-abroad group (SA, N = 47).

Task	Measure	Group	Pre-test	Post-test
			Mean (SD)	*Mean (SD)*
Lexical access task	Accuracy	AH-regular	.83 (.06)	.83 (.05)
		AH-intensive	.87 (.05)	.88 (.05)
		SA	.88 (.04)	.89 (.04)
	RT	AH-regular	1284 (291)	1038 (194)
		AH-intensive	1188 (265)	1080 (223)
		SA	1137 (250)	981 (190)
	CV	AH-regular	.36 (.09)	.31 (.09)
		AH-intensive	.34 (.09)	.33 (.08)
		SA	.33 (.08)	.29 (.09)
Grammatical processing task	Accuracy	AH-regular	.83 (.08)	.85 (.08)
		AH-intensive	.86 (.06)	.88 (.06)
		SA	.86 (.07)	.88 (.05)
	RT	AH-regular	3375 (452)	3008 (458)
		AH-intensive	3248 (436)	2936 (371)
		SA	3289 (424)	2871 (379)
	CV	AH-regular	.38 (.05)	.38 (.05)
		AH-intensive	.39 (.06)	.39 (.05)
		SA	.38 (.05)	.39 (.06)
Semantic processing task	Accuracy	AH-regular	.84 (.12)	.87 (.11)
		AH-intensive	.93 (.05)	.95 (.04)
		SA	.93 (.07)	.94 (.05)
	RT	AH-regular	3596 (669)	3061 (567)

		AH-intensive	3011 (347)	2737 (308)
		SA	3049 (447)	2691 (441)
	CV	AH-regular	.29 (.06)	.25 (.05)
		AH-intensive	.28 (.06)	.26 (.05)
		SA	.26 (.05)	.23 (.06)
Vocabulary size test	Vocabulary size score	AH-regular	116 (28)	121 (27)
		AH-intensive	148 (17)	149 (17)
		SA	132 (23)	139 (22)

Note: The values of Accuracy are proportions and RT values are in milliseconds.

3.3.1 Vocabulary size

As Table 3.3 and Figure 3.1 show, the AH-intensive group outperformed the SA group (*β* = 15.97, *SE* = 4.69, *t* = 3.40, *p* < .001), who in turn outperformed the AH-regular group (*β* = -15.12, *SE* = 4.58, *t* = -3.30, *p* < .001) in terms of vocabulary in the *pre-test*. In general, participants' vocabulary size in the post-test was significantly larger than in the pre-test (*β* = 4.56, *SE* = 1.11, *t* = 4.10, *p* < .001). The SA group made more progress than the AH-intensive group (*β* = -6.09, *SE* = 2.77, *t* = -2.20, *p* = .030), but there was no significant difference between vocabulary improvement of the SA and the AH-regular groups (*β* = -2.58, *SE* = 2.71, *t* = -0.95, *p* = .342). Furthermore, we split the data by group (see Appendix Table D1) and found that both the SA group (β = 7.45, SE = 1.80, t = 4.14, p < .001) and the AH-regular group (*β* = 4.87, *SE* = 2.29, *t* = 2.13, *p* = .038) made significant improvement in vocabulary over time but the AH-intensive group (*β* = 1.35, *SE* = 1.49, *t* = 0.91, *p* = .369) did not.

Table 3.3: Estimates of performance of participant groups in the vocabulary size test

	B	***SE***	***t***	***p***
(Intercept)	131.86	1.88	70.09	< .001
TimePostvsPre	4.56	1.11	4.10	< .001
GroupAH-regularvsSA	-15.12	4.58	-3.30	< .001
GroupAH-intensivevsSA	15.97	4.69	3.40	< .001
TimePostvsPre:GroupAH-regularvsSA	-2.58	2.71	-0.95	0.342
TimePostvsPre:GroupAH-intensivevsSA	-6.09	2.77	-2.20	0.030

*Note: Model specification: lmer(Vocab ~ Time*Group + (1|Subject)).*

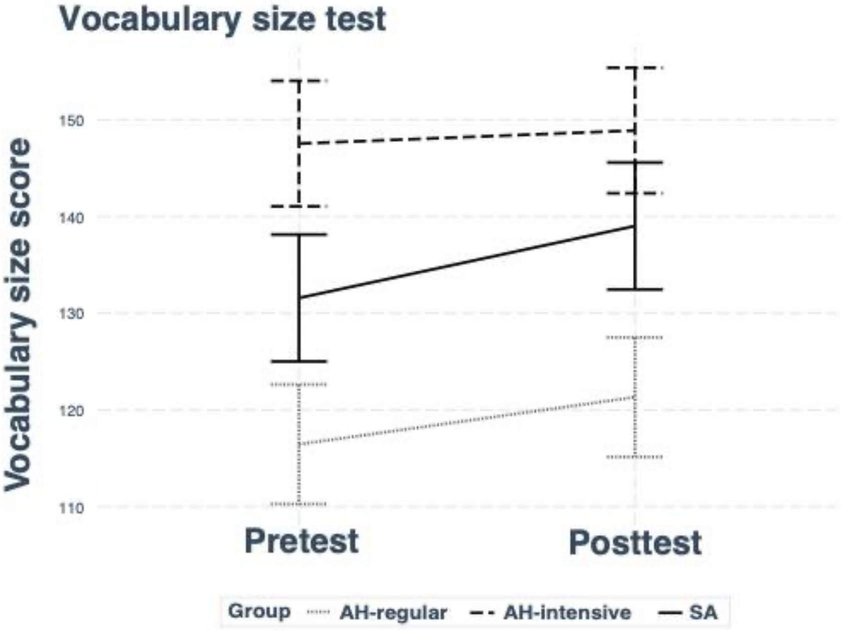

Figure 3.1: Interaction between Time and Group effects in the vocabulary size model. Y-axis does not start from zero. Error bars represent standard errors.

3.3.2 Processing accuracy

As shown in Table 3.4 and Figure 3.2, the SA group outperformed the AH-regular group (*β* = -0.80, *SE* = 0.13, *z* = -6.37, *p* < .001), but did not differ from the AH-intensive group (*β* = -0.16, *SE* = 0.13, *z* = -1.19, *p* = .233) in terms of processing accuracy in the *pre-test*. Accuracy performance in the pre-test did not differ across tasks (*β* = -0.52, *SE* = 0.30, *z* = -1.73, *p* = .084; *β* = 0.10, *SE* = 0.31, *z* = 0.32, *p* = .748). However, the difference between the AH-regular and SA groups in the lexical access task was larger than that in the grammatical processing task (*β* = 0.44, *SE* = 0.17, *z* = 2.55, *p* = .011), but was slightly smaller than that in the semantic processing task (*β* = -0.40, *SE* = 0.21, *z* = -1.94, *p* = .053).

As for the accuracy difference between pre-test and post-test, participants had higher accuracy in the post-test than in the pre-test across tasks (*β* = 0.23, *SE* = 0.04, *z* = 5.15, *p* < .001). The general accuracy improvement did not differ across either task (*β* = 0.15, *SE* = 0.10, *z* = 1.46, *p* = .146; *β* = 0.09, *SE* = 0.11, *z* = 0.80, *p* = .424) or group (*β* =

-0.09, *SE* = 0.08, *z* = -1.11, *p* = .269; *β* = 0.05, *SE* = 0.09, *z* = 0.58, *p* = .565). None of the three-way interactions was significant.

Table 3.4: Fixed-effect estimates of accuracy performance of participant groups in the three processing tasks.

	B	***SE***	*z*	*p*
(Intercept)	2.74	0.13	20.56	< .001
TimePostvsPre	0.23	0.04	5.15	< .001
GroupAH-regularvsSA	-0.80	0.13	-6.37	< .001
GroupAH-intensivevsSA	-0.16	0.13	-1.19	.233
Task2vs1	-0.52	0.30	-1.73	.084
Task3vs1	0.10	0.31	0.32	.748
TimePostvsPre:GroupAH-regularvsSA	-0.09	0.08	-1.11	.269
TimePostvsPre:GroupAH-intensivevsSA	0.05	0.09	0.58	.565
TimePostvsPre:Task2vs1	0.15	0.10	1.46	.146
TimePostvsPre:Task3vs1	0.09	0.11	0.80	.424
GroupAH-regularvsSA:Task2vs1	0.44	0.17	2.55	.011
GroupAH-intensivevsSA:Task2vs1	0.24	0.18	1.34	.180
GroupAH-regularvsSA:Task3vs1	-0.40	0.21	-1.94	.053
GroupAH-intensivevsSA:Task3vs1	0.19	0.22	0.88	.381
TimePostvsPre:GroupAH-regularvsSA:Task2vs1	0.18	0.18	0.98	.326
TimePostvsPre:GroupAH-intensivevsSA:Task2vs1	0.07	0.19	0.34	.736
TimePostvsPre:GroupAH-regularvsSA:Task3vs1	0.33	0.21	1.56	.118
TimePostvsPre:GroupAH-intensivevsSA:Task3vs1	0.24	0.23	1.04	.300

*Notes: 1. Model specification: glmer(Accuracy ~ Time*Group*Task + (1 + Time + Group|Item_number) + (1 +Time + Task|SubjectNo)).*

2. Tasks 1, 2 and 3 refer to the lexical access, grammatical processing and semantic processing task, respectively.

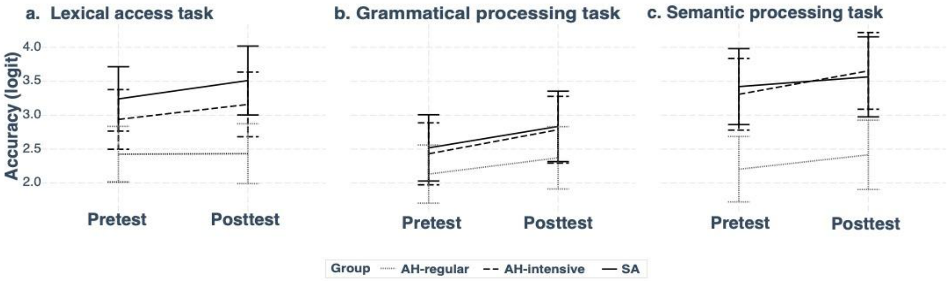

Figure 3.2: Interaction between Time, Group and Task effects in the processing accuracy model with performance on different linguistic tasks in different panels. Y-axis does not start from zero. Error bars represent standard errors.

3.3.3 Processing speed

As Table 3.5 and Figure 3.3 show, at the *pre-test*, similar to the accuracy model, the SA group outperformed the AH-regular group ($\beta = 0.10$, $SE = 0.02$, $t = 4.30$, $p < .001$), but did not differ significantly from the AH-intensive group ($\beta = 0.00$, $SE = 0.03$, $t = 0.16$, $p = .876$) in terms of processing speed. RT performance at the pre-test differed across tasks ($\beta = 0.29$, $SE = 0.08$, $t = 3.58$, $p < .001$; $\beta = 0.21$, $SE = 0.09$, $t = 2.39$, $p = .018$), which was expected as the stimuli duration differs across the tasks. Furthermore, the difference between the AH-regular and SA groups in the lexical access task was also larger than that in the grammatical processing task ($\beta = -0.08$, $SE = 0.03$, $t = -3.15$, $p = .002$) and did not differ significantly from that of the semantic processing task ($\beta = 0.04$, $SE = 0.03$, $t = 1.36$, $p = .174$). Similarly, the difference between the AH-intensive and SA group in the lexical access task was larger than that in the grammatical processing task ($\beta = -0.06$, $SE = 0.03$, $t = -2.17$, $p = .031$), but did not differ significantly from that of the semantic processing task ($\beta = -0.06$, $SE = 0.03$, $t = -1.95$, $p = .053$).

As for the RT difference between pre-test and post-test, participants generally responded faster in the post-test than in the pre-test ($\beta = -0.12$, $SE = 0.00$, $t = -49.89$, $p < .001$). The progress that participants made in RT did not differ significantly across tasks ($\beta = 0.01$, $SE = 0.01$, $t = 1.70$, $p = .089$; $\beta = 0.01$, $SE = 0.01$, $t = 1.48$, $p = .139$). The AH-regular group made more progress in processing speed than the SA group ($\beta = -0.03$, $SE = 0.01$, $t = -4.55$, $p < .001$), who in turn made more progress than the AH-intensive group ($\beta = 0.03$, $SE = 0.01$, $t = 4.98$, $p < .001$). The degree of progress made by the groups in RT was affected by tasks. More specifically, the difference between the SA and the AH-regular group, in terms of progress in RT performance, in the lexical access task was significantly different from that in the grammatical processing (*β* = 0.09, *SE* = 0.01, *t* = 6.55, *p* < .001) and semantic processing tasks (*β* = 0.04, *SE* = 0.01, *t* = 2.94, *p* < .01). To clarify the three-way interactions between Time, Group and Task effects, we split the dataset by task and then fitted models for each task dataset (see Appendix Table D2). The AH-regular group made more progress than the SA group in the lexical access task (*β* = -0.07, *SE* = 0.01, *t* = -7.02, *p* < .001), which was also observed in the semantic processing task but with a smaller effect size (*β* = -0.03, *SE* = 0.01, *t* = -3.04, *p* = .002). However, the reverse pattern for the same group comparison was found in the grammatical processing task (*β* = 0.02, *SE* = 0.01, *t* = 1.97, *p* = .049).

Table 3.5: Fixed-effect estimates of RT performance of participant groups in the three processing tasks.

	B	***SE***	***t***	***p***
(Intercept)	3.20	0.45	7.07	< .001
TimePostvsPre	-0.12	0.00	-49.89	< .001
GroupAH-regularvsSA	0.10	0.02	4.30	< .001
GroupAH-intensivevsSA	0.00	0.03	0.16	.876
Task2vs1	0.29	0.08	3.58	< .001
Task3vs1	0.21	0.09	2.39	.018
log(audio_duration)	0.64	0.06	10.03	< .001
Trial_number	-0.00	0.00	-11.03	< .001
TimePostvsPre:GroupAH-regularvsSA	-0.03	0.01	-4.55	< .001
TimePostvsPre:GroupAH-intensivevsSA	0.03	0.01	4.98	< .001
TimePostvsPre:Task2vs1	0.01	0.01	1.70	.089
TimePostvsPre:Task3vs1	0.01	0.01	1.48	.139
GroupAH-regularvsSA:Task2vs1	-0.08	0.03	-3.15	.002
GroupAH-intensivevsSA:Task2vs1	-0.06	0.03	-2.17	.031
GroupAH-regularvsSA:Task3vs1	0.04	0.03	1.36	.174
GroupAH-intensivevsSA:Task3vs1	-0.06	0.03	-1.95	.053
TimePostvsPre:GroupAH-regularvsSA:Task2vs1	0.09	0.01	6.55	< .001
TimePostvsPre:GroupAH-intensivevsSA:Task2vs1	0.02	0.01	1.28	.201
TimePostvsPre:GroupAH-regularvsSA:Task3vs1	0.04	0.01	2.94	.003
TimePostvsPre:GroupAH-intensivevsSA:Task3vs1	0.01	0.01	0.45	.655

*Notes: 1. Model specification: lmer(log_RT ~ Time*Group*Task + log_audio_duration + Trial_number + (1 + Task|SubjectNo) + (1+Group|Item_number)).*

2. *Tasks 1, 2 and 3 refer to the lexical access, grammatical processing and semantic processing task, respectively.*

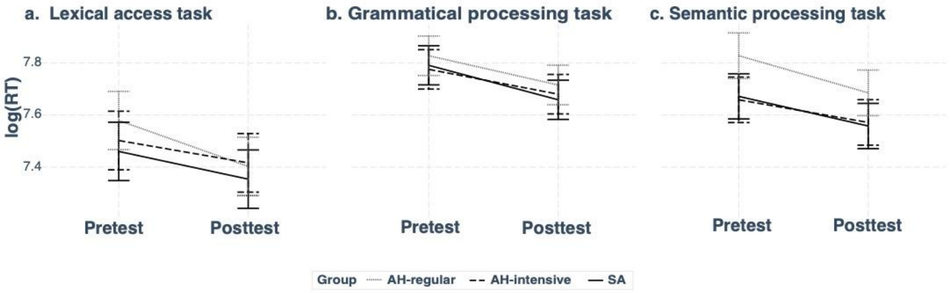

Figure 3.3: Interaction between Time, Group and Task effects in the processing speed model with performance on different linguistic tasks in different panels. Y-axis does not start from zero. Error bars represent standard errors.

3.3.4 Processing stability

As is shown in Table 3.6 and Figure 3.4, the three participant groups did not differ significantly in processing stability in the *pre-test* (β = 0.01, *SE* = 0.01, *t* = 1.56, *p* = .121; β = 0.01, *SE* = 0.01, *t* = 1.29, *p* = .198). RTs in the lexical access task were more stable than those in the grammatical processing task (β = 0.04, *SE* = 0.01, *t* = 6.40, *p* < .001), and less stable than those in the semantic processing task (β = -0.07, *SE* = 0.01, *t* = -10.78, *p* < .001). Moreover, the difference between the AH-regular and SA groups on the lexical access task was significantly different from that on the grammatical processing task (β = -0.03, *SE* = 0.02, *t* = -2.09, *p* = .037), but did not differ significantly from that of the semantic processing task (β = 0.00, *SE* = 0.02, *t* = 0.27, *p* = .789).

As for the CV difference between pre-test and post-test, participants' RTs were generally more stable in the post-test than in the pre-test (β = -0.02, *SE* = 0.00, *t* = -5.30, *p* < .001). The progress participants made in the lexical access task did not differ significantly from that in the semantic processing task (β = 0.01, *SE* = 0.01, *t* = 0.64, *p* = .526), but was larger than that in the grammatical processing (β = 0.04, *SE* = 0.01, *t* = 4.02, *p* < .001). The stability difference between the pre-test and post-test was not modulated by group (β = -0.00, *SE* = 0.01, *t* = -0.18, *p* = .854; β = 0.01, *SE* = 0.01, *t* = 1.00, *p* = .317). This model had no significant three-way interaction between Time, Group and Task effects.

Table 3.6: Fixed-effect estimates of CV performance of the participant groups in the three processing tasks.

	B	*SE*	*t*	*p*
(Intercept)	0.33	0.00	91.41	< .001
TimePostvsPre	-0.02	0.00	-5.30	< .001
GroupAH-regularvsSA	0.01	0.01	1.56	.121
GroupAH-intensivevsSA	0.01	0.01	1.29	.198
Task2vs1	0.04	0.01	6.40	< .001
Task3vs1	-0.07	0.01	-10.78	< .001
TimePostvsPre:GroupAH-regularvsSA	-0.00	0.01	-0.18	.854
TimePostvsPre:GroupAH-intensivevsSA	0.01	0.01	1.00	.317
TimePostvsPre:Task2vs1	0.04	0.01	4.02	< .001
TimePostvsPre:Task3vs1	0.01	0.01	0.64	.526
GroupAH-regularvsSA:Task2vs1	-0.03	0.02	-2.09	.037
GroupAH-intensivevsSA:Task2vs1	0.00	0.02	0.27	.790
GroupAH-regularvsSA:Task3vs1	0.00	0.02	0.27	.789
GroupAH-intensivevsSA:Task3vs1	0.02	0.02	1.01	.313
TimePostvsPre:GroupAH-regularvsSA:Task2vs1	0.00	0.02	0.21	.835
TimePostvsPre:GroupAH-intensivevsSA:Task2vs1	-0.04	0.02	-1.85	.065
TimePostvsPre:GroupAH-regularvsSA:Task3vs1	0.00	0.02	0.15	.882
TimePostvsPre:GroupAH-intensivevsSA:Task3vs1	-0.02	0.02	-1.04	.298

Notes: 1. Model specification: lmer(CV_pp ~Time Group*Task +(1 + Time|SubjectNo)).*

2. Tasks 1, 2, and 3 refer to the lexical access, grammatical processing, and semantic processing task, respectively.

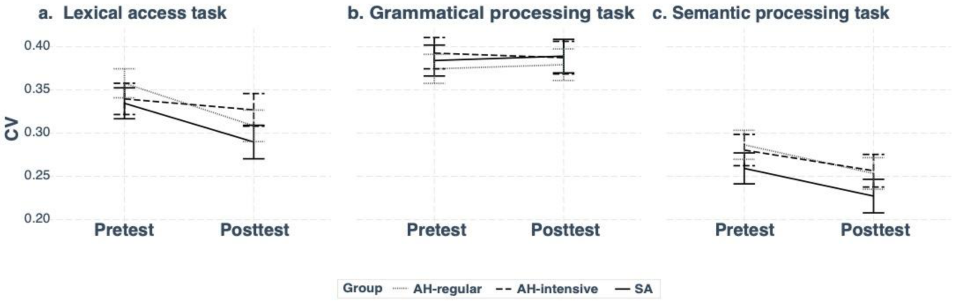

Figure 3.4: Interaction between Time, Group and Task effects in the processing stability model with performance on different linguistic tasks in different panels. Y-axis does not start from zero. Error bars represent standard errors.

In summary, with regard to baseline proficiency, the AH-intensive group had the largest vocabulary size at the pre-test, followed by the SA group who in turn outperformed the AH-regular group. Moreover, the AH-intensive and SA groups did not differ from each other but outperformed the AH-regular group in terms of processing efficiency at the pre-test. As for the performance difference between the pre-test and the post-test, the SA and AH-regular groups made comparable progress in vocabulary size, but the AH-intensive group did not make any significant progress. Meanwhile, the AH-regular group made more progress in processing speed than the SA groups, who made more progress than the AH-intensive group. However, the progress in accuracy and stability of processing was not significantly different among these three learner groups.

3.4 Discussion

3.4.1 The effect of learning contexts on vocabulary size

Our first research question was whether the improvement in auditory vocabulary over the course of an academic year is conditioned by learning context. We found that the SA and AH-regular groups made comparable progress in vocabulary size, and their progress was larger than that of the AH-intensive group. Post-hoc analyses on vocabulary improvement in each of the separate groups showed that the AH-intensive group did not make significant progress but both the SA and AH-regular groups did. These results can be broken down into two SA-AH comparisons. First, although the SA group made significantly more progress than the AH-intensive group, the significance may be driven either by the fact that the AH-intensive group did not improve, or by the advantageous effects of the SA context on vocabulary acquisition as compared to the AH context. Note that even at post-test, the mean vocabulary scores of the AH-intensive group were still numerically higher than those of the SA group. Thus, it cannot be concluded that the SA context was more effective in facilitating vocabulary acquisition, based on the comparison between the SA and the AH-intensive groups. Second, vocabulary improvement of the SA group over an academic year abroad did not differ significantly from that of the AH-regular group. However, it is unclear whether this similar improvement pattern for these two groups should be attributed (partly) to differences in baseline vocabulary size.

Taken together, these results do not provide clear evidence to support our initial hypothesis about the facilitative role of the SA context in vocabulary acquisition relative to the AH context. Therefore, even though adult L2 learners may benefit more from the extra incidental learning opportunities provided by SA context relative to the AH context, the magnitude of the learning-context advantage on vocabulary improvement seems small or non-existent. One possible reason is that SA learners may have faced problems with social integration, leading to a low degree of immersion (see Coleman, 2013 for a model of socialization while abroad). That is, the supposedly rich input, output, and interaction in the SA context may turn out to be shallow due to integration problems, which may make learning advanced vocabulary difficult. On the other hand, it is conceivable that the SA group may have acquired more vocabulary used in their specific discipline (e.g., architecture, chemistry of philosophy) during class and more vocabulary relevant to their immediate living experience (e.g., names of grocery items or cooking utensils) out of class than their at-home peers, but such vocabulary gains may not be effectively detected by the PPVT™-4.

The present study complements studies on the impact of learning contexts on vocabulary acquisition in reading, writing and speaking activities (e.g., Briggs, 2015; DeKeyser, 1991; Dewey, 2004, 2008; Ife, Vives Boix, & Meara, 2000; Llanes & Muñoz, 2009; Milton & Meara, 1995). These studies also present a mixed picture of SA effects, especially on reading vocabulary development. More specifically, some studies reported substantially greater gains in reading vocabulary in SA contexts compared to AH contexts (e.g., Milton & Meara, 1995), whereas others found no significant difference between intensive domestic instruction and study-abroad contexts (e.g., Dewey, 2004, 2008; Serrano, Llanes, & Tragant, 2011). As for productive vocabulary used in speaking and writing, no significant advantage of SA over AH contexts was found concerning the acquisition of new words (Collentine, 2004; Freed, So and Lazar, 2003).

In contrast to Dewey (2008) and Serrano et al. (2011), who reported comparable vocabulary development in intensive domestic immersion and SA contexts, the improvement patterns of the AH-intensive and SA groups in the present study seem to suggest that the intensive domestic program may not be effective in enhancing listening vocabulary for relatively advanced learners. In other words, the fact that there was no significant difference in vocabulary size of the AH-intensive group between pre-test and post-test could mean that this group had reached a plateau in auditory vocabulary acquisition. However, since the average vocabulary score of the AH-intensive group is far from the ceiling performance according to the PPVT™-4 manual, there should be plenty of room for vocabulary growth. Therefore, the lack of improvement of auditory vocabulary in the AH context is unexpected. Previous studies (e.g., Han, 2013; Han & Odlin, 2006; Selinker, 1972) have speculated on a stabilization or fossilization phenomenon where L2 language proficiency stops improving regardless of abundant target-language exposure. However, these studies have mostly been carried out in study/residence abroad contexts where learners are immersed in their L2. We speculate that learners could also reach a point of stabilization in auditory vocabulary acquisition in AH contexts with generally impoverished target language exposure, but note that measurement at two time points does not allow a firm conclusion about stabilization.

3.4.2 The effect of learning contexts on processing efficiency

Our second research question addressed the question of whether the SA and the AH learning contexts differed in facilitating the development of L2 listening processing skills. Participant groups all improved in terms of accuracy and stability of processing, but there were no significant group differences in the amount of improvement for these two measures. As for the speed of processing, the SA group made more progress than the AH-intensive group and less improvement than the AH-regular group. We interpret these results as follows. Firstly, the comparison between the SA group and the AH-intensive group suggests that the SA learners improved their speed of processing more rapidly than their at-home peers. As processing efficiency of these two groups did not differ at the pre-test, it is likely the observed effects are not due to baseline confounds, but rather should be explained by the effect of learning context. Secondly, the fact that the SA group showed less improvement than the AH-regular group in the speed of processing seems to contradict our hypothesis that the SA context would facilitate processing efficiency better than the AH context. However, steep learning curves for low-proficiency participants in reaction time were commonly observed in previous studies (e.g., van den Bosch, Segers & Verhoeven, 2019). Since the AH-regular group had lower proficiency than the AH-intensive and the SA group at pre-test, it can be argued that the AH-regular group improved relatively fast due to their lower starting level at the beginning of the academic year. Therefore, the effect of learning context, substantiated by the comparison between SA and AH-regular groups, should be considered in light of baseline proficiency.

Therefore, provided equal starting levels, studying abroad is more beneficial for enhancing L2 processing speed over remaining at home. This agrees with our initial hypothesis about the facilitative effect of SA learning contexts on processing efficiency. The findings of the present study complement, though not always align with, those of previous studies on fluency (e.g., Freed, Segalowitz & Dewey, 2004; Sasaki, 2007; Segalowitz & Freed, 2004), another aspect of language processing. Previous studies have usually associated greater gains in fluency with SA contexts relative to AH contexts. Freed, Segalowitz and Dewey (2004) compared oral fluency development (L1 English, L2 French) in AH, SA and domestic immersion contexts over one semester, and found that the SA group improved more than the AH group but less than the domestic immersion group. Note, however, that the questionnaires of that study revealed that the domestic immersion group surprisingly used the target L2 French more than the SA group. In contrast, Sasaki (2007) compared changes in the writing of SA learners and domestic English majors (L1 Japanese, L2 English) during one year, and found that the SA group improved their English writing fluency but the AH group did not. The present study provides evidence that similar to fluency development in oral and written production, the speed of L2 processing in listening comprehension develops more rapidly in an SA context than an AH context. This suggests that the development of L2 processing skills is subject to the effect of learning contexts.

Furthermore, the advantage of the SA group over the AH-intensive group in facilitating the development of L2 processing efficiency was not constrained to specific linguistic processes. This means that the effect of learning contexts on improving L2 processing efficiency manifests in all three linguistic processes studied (i.e., lexical access, grammatical processing, and semantic interpretation). Previous studies on the effect of learning context on either lexical access or grammatical processing, however, seem to present a different picture. Segalowitz and Freed (2004) tested SA and AH learners with a semantic classification task in a pre-test and a post-test, and found no differential gains in lexical access as a function of learning contexts. Similarly, Isabelli-García (2010) and Pliatsikas and Marinis (2013a) also reported no difference between these two contexts in grammar acquisition, such as the acquisition of gender agreement and past tense. The present study contradicts with these studies in that we found learning-context effects on language processing at lexical, morphosyntactic, and semantic levels. However, the abovementioned studies all measured L2 processing in reading, while the present study targeted that in listening. The inconsistent findings between the previous studies and the present study suggest that the SA context may have differential effects on L2 processing in listening and reading activities. This may be related to the fact that formal instruction in AH contexts usually focuses on reading rather than listening. Thus, when learners move to an SA context, their L2 listening ability is likely to develop more rapidly than their reading. In addition, comparing task performance at the pre-test and the post-test, we found the grammatical processing task to show the least difference between the participant groups. This aligns with and reconfirms Chapter 2's conclusion that, compared to lexical access and semantic processing, grammar sensitivity is the area that least distinguished the proficiency of nonnative groups. However, though between-group differences in L2 processing were enlarged for specific linguistic tasks at both testing times, performance on the different linguistic tasks developed similarly over time. One possible reason is that the improvement in L2 processing over one academic year may not be big enough to show the finer differences in the development of linguistics processes, or that the measurement used in the present study may not be sensitive enough to show such differences.

3.4.3 Implications, limitations and future directions

This study has pedagogical implications for language learning in both SA and AH contexts.

1) Compared to the domestic non-English-major postgraduates and study-abroad learners, the domestic English-major postgraduates made the least progress in both vocabulary knowledge and processing efficiency. Therefore, the AH context seems to be less effective for L2 listening development of relatively high-proficiency learners. The curriculum of domestic English majors, especially for postgraduates, may need to incorporate more opportunities for interactive language practice. Otherwise, a sojourn abroad might be conducive to further progress.
2) The fact that the SA group did not differ from the AH-regular group in vocabulary improvement supports the view that the role of incidental vocabulary learning may be limited. SA learners ought to seek systematic vocabulary learning activities to achieve greater learning outcomes.
3) We recommend including language processing tests in L2 learning and teaching practice. Such tests offer an accurate measurement of L2 processing efficiency which reveals learners' L2 strengths and weaknesses from a skill acquisition perspective, thus providing learners, teachers, and researchers more insight on SLA.

This study has certain limitations. Firstly, vocabulary knowledge of participant groups differed from each other at the pre-test, which may confound the effect of learning contexts to some degree. As it is difficult to manipulate baseline proficiency levels of students in natural learning contexts, previous studies often tend to ignore the effect of baseline proficiency level and directly compare SA and AH learners. Our study addresses this problem by using proper statistical analysis methods (i.e., LMMs) and taking baseline proficiency differences into account when interpreting our results. Note that these analytical methods are a statistical control of baseline proficiency differences, but do not account for other potential differences between the groups. To investigate whether our vocabulary developmental pattern results should indeed be attributed to baseline proficiency differences, replication with participant groups matched on baseline vocabulary knowledge is recommended. Secondly, it should be noted that performance on the processing tasks is sensitive to repetition such that the general progress participants made may be influenced by test-retest effects. However, it is the relative between-group difference in progress that is the major concern of this study, instead of the absolute amount of progress.

To our knowledge, this is the first study to examine the effect of SA and AH learning contexts on the development of L2 listening proficiency from both language knowledge and language processing perspective. Future studies are encouraged to further this investigation in other learning contexts, such as heritage language learning, residence abroad, domestic immersion, and computer-assisted learning contexts, to shed more light on the relation between L2 listening development and learning contexts. Moreover, the linguistic benefits of the SA context may be affected by the duration of an SA experience. The effect of study-abroad duration, especially the contrast between long- and short-term study abroad, should be addressed by future research.

3.5 Conclusions

We found that, in terms of vocabulary gains, the SA group made more progress than the AH-intensive group, but did not differ significantly from the AH-regular group. However, it is difficult to tell whether this differential vocabulary growth should be attributed to the learning context, or may be due to different starting levels. At the same time, the SA group made more progress than the AH-intensive group but less progress than the AH-regular group in speed of lexical, morphosyntactic, and semantic processing. Since the SA and AH-intensive groups started off with equal processing efficiency levels at pre-test while the SA and AH-regular groups did not, we argue that the difference between the SA and AH-intensive groups in terms of processing efficiency improvement should be attributed to the effect of learning context. More specifically, this suggests that the SA context facilitates the acquisition of processing skills (processing speed in particular) better than the AH context. To sum up, this study demonstrates that study abroad is an effective intervention for developing L2 processing efficiency with less clear effects on vocabulary acquisition.

Chapter 4: Individual differences and L2 listening in study-abroad and at-home learning contexts

Abstract

This study examines the association between several individual-difference factors and L2 listening comprehension in study-abroad (SA) and at-home (AH) learning contexts, investigating the possible interplay between individual-difference factors and learning context. 143 participants from SA, AH-regular, and AH-intensive contexts took a battery of L2 listening tests twice, with an interval of one academic year. These tests specifically measured auditory vocabulary size and spoken-language processing efficiency (i.e., accuracy, speed, and stability of processing). Tests and questionnaires measuring individual differences in language aptitude, working memory, mental well-being, language exposure, and social interaction were administered once. Mixed-effects regression modelling showed that none of these individual-difference factors predicted participants' improvement over this academic year, but some were nevertheless stable predictors of listening proficiency at both pre- and post-test. Language aptitude was the only individual-difference factor that predicted vocabulary size, and it did so across learning contexts. Both language aptitude and exposure correlated with accuracy and speed of processing across learning contexts. Working memory, mental well-being, and social interaction did not relate to any of the listening measures. Importantly, as no interactions between learning context and individual-difference factors were observed, second language learning seems to be similarly related to individual capacities across different learning contexts.

4.1 Introduction

Second language acquisition is known for its great variability in acquisition rate and ultimate attainment among L2 learners (see e.g., Dewaele, 2009; Tagarelli, Ruiz, Vega, & Rebuschat, 2016). This variability may be explained by differences in learning context as well as individual differences in cognitive, affective, social, and linguistic factors (for reviews of individual differences in L2 learning, see e.g., Dewaele, 2009; Skehan, 2014). Since learning context and individual-difference factors jointly affect the various aspects of L2 learning, understanding variability in L2 development requires joint investigation of these factors with careful consideration of their potential interplay (see Dörnyei, 2009; DeKeyser, 2012; Faretta-Stutenberg & Morgan-Short, 2018; Robinson, 2001; Sanz, 2005). Indeed, previous studies have suggested that individual-difference factors in cognitive abilities interact with learning context in mediating second language learning processes and outcomes (e.g., Sunderman & Kroll, 2009; Tokowicz, Michael, & Kroll, 2004; Faretta-Stutenberg & Morgan-Short, 2018). For instance, Faretta-Stutenberg and Morgan-Short (2018) reported that gains in sensitivity to L2 syntax were associated with individual differences in procedural learning ability and working memory for study-abroad (SA) learners but not for at-home (AH) learners. However, only a few studies have studied the interplay between learning context, individual differences, and L2 learning. To the best of our knowledge, none of these few studies have addressed the development of L2 listening proficiency and its components. To fill this gap, the present study used a pre- and post-test design, testing Chinese learners of English at the beginning and end of an academic year, to examine the effect of several individual-difference factors in three learning contexts and to investigate the possible interplay between learning context, individual-difference factors, and L2 listening development.

4.1.1 Learning context

L2 learning contexts (e.g., regular classroom, study abroad, or domestic immersion programs) vary widely in terms of quality and quantity of input and opportunities for interaction, each of which may contribute to the diversity of L2 development trajectories. The SA context is commonly believed to be the best L2 learning context as it is supposed to provide learners with native input, informative feedback, and ample possibility for interaction. The AH context, on the other hand, is often characterized by inadequate L2 exposure (albeit with access to authentic L2 input via digital platforms), overreliance on rote learning, and rare opportunities for real L2 communication. Studies that compare SA and AH contexts in terms of fostering language development have focused on various aspects of language proficiency, such as general language proficiency, writing proficiency, vocabulary size, and oral fluency. Some studies found that SA learners had greater gains than AH learners in e.g., conventionalized word combinations (Foster, Bolibaugh, & Kotula, 2014), grammar (Marqués-Pascual, 2011; Faretta-Stutenberg & Morgan-Short, 2018), native-like accent (Muñoz & Llanes, 2014), pragmatic performance (Matsumura, 2001), writing proficiency (Sasaki, 2011), and oral proficiency (Segalowitz & Freed, 2004). Others, however, reported marginal or no differences between SA and AH learners in e.g., grammar (Isabelli-Garcia, 2010; Håkansson & Norrby, 2010), receptive vocabulary (Chapter 3), and general writing proficiency (Knoch, Rouhshda, Oon, & Storch, 2015).

Each learning context may have its own advantages and limitations when it comes to facilitating specific aspects of L2 acquisition. Learners' oral fluency, measured by speech rate (i.e., words per minute) and fluent run length (i.e., number of words in fluent speech runs not containing any silent or filled dysfluencies), usually improves after studying abroad (see Mora & Valls-Ferrer, 2012; Segalowitz & Freed, 2004; Serrano, Tragant, & Llanes, 2011). However, SA learners may not necessarily outperform AH learners in terms of accuracy and complexity of oral production, such as frequency of errors, length and syntactic complexity of sentences, as evidenced by mixed results across studies (e.g., Håkansson & Norrby, 2010; Llanes & Muñoz, 2013; Marqués-Pascual, 2011; Collentine, 2004). Similarly, for written production, benefits of studying abroad have usually been reported for writing fluency but not necessarily for accuracy or complexity (see Knoch, Rouhshda, Oon, & Storch, 2015; Sasaki, 2007). Thus, studying abroad does not seem to benefit all aspects of language learning to the same extent.

4.1.2. Individual differences

L2 learners may acquire language at different paces due to individual differences in cognitive, emotional, social, and linguistic factors, even when they are learning a language in exactly the same environment. To unravel the complex relationship between language learning and individual differences, we will first review the literature on five individual-difference factors that have been associated with language learning (i.e., language aptitude, working memory, mental well-being, language exposure, and social interaction), followed by a discussion of the potential interaction between these factors and learning context.

Language aptitude and working memory

Language aptitude, or the talent to learn new languages, is an important individual-difference factor that can be defined as a combination of cognitive and perceptual abilities that are advantageous to second language acquisition (Granena, 2013). However, there is no consensus on how to operationalize language aptitude. The two most widely recognized language aptitude tests, the Modern Language Aptitude Test (MLAT; Carroll & Sapon, 1959) and the LLAMA Language Aptitude Tests (Meara, 2005), break down language aptitude into different components (e.g., associative memory, phonemic coding, grammatical sensitivity), arguing that these are involved in language acquisition. Working memory has later been included by some recent literature as a component of language aptitude as well (e.g., DeKeyser & Koeth, 2011; Skehan, 2012). The cognitive skill of working memory reflects the dual components of mental operations: information storage and processing, and is associated with language aptitude (Li, 2016; Linck, Osthus, Koeth, & Bunting, 2014).

Language aptitude has often been found to be related to L2 proficiency (see Li, 2016 for a review), but the role of aptitude may vary depending on learning phase. Aptitude tests (e.g., MLAT) predict language learning during the initial phase, but may not predict language attainment well for advanced learners. This could be because different components of aptitude, or distinct combinations of abilities, are at play at the beginning and the advanced stages of language learning, as speculated by Carroll (1990). For example, the role of pragmatic abilities may be marginal for beginners learning an L2 in classroom settings but will gain importance once learners are advanced enough to be involved in real communication. While previous aptitude studies have generally focused on the initial stages of language learning, the present study investigates whether aptitude differences still play a role amongst intermediate to advanced learners.

Studies that have investigated whether aptitude is drawn upon differently in different (experimentally-manipulated) learning conditions show mixed results (e.g., Farshi & Tavakoli, 2019; Robinson, 2002, 2005; Reber, Walkenfield, & Hernstadt, 1991; Carroll, 1990). Some studies (e.g., Farshi & Tavakoli, 2019; Robinson, 1997; Sheen, 2007) found that aptitude was related to pace of learning in both implicit-learning conditions (without conscious awareness of rules to be mastered) and explicit-learning conditions (involving explicit explanation of metalinguistic knowledge and rules). These authors argued that L2 learning for adults was associated with aptitude, regardless of the conditions in which learning occurred. Other studies (e.g., Granena, 2016; Krashen, 1985), however, argued that implicit learning was not related to aptitude differences reflecting conscious L2 learning. As aptitude tests have been criticized for having low predictive power for implicit learning (see Li, 2016), Robinson (2005) argued that aptitude tests should be supplemented by other measures like working memory. Learners with better working memory are better equipped to pay attention simultaneously to meaning and form, especially under learning conditions with no intentional focus on form (Robinson, 2002). Hence, aptitude and working memory may each account for unique variance in L2 learning, which is why we will keep them separate.

Mental well-being

Mental well-being is suggested to play an important role in language acquisition (MacIntyre,Gregersen, & Mercer, 2019). Several empirical studies have shown detrimental effects of negative emotions (e.g., anxiety and stress) on L2 learning, e.g., through triggering communicative breakdown (Dewaele, 2002) and lowered willingness to speak (MacIntyre, Baker, Clément, & Donovan, 2003). Whereas negative emotions may have a detrimental effect on language learning, Fredrickson's (2001) broaden-and-build theory argued that positive emotions, such as joy and interest, broaden and then build upon personal and social resources. Fredrickson and Branigan (2005) showed that positive emotions were related to being flexible and open to information, and, conversely, negative emotions were related to narrow attentional focus. MacIntyre and Gregersen (2012) further applied this to language learning and posited that positive emotions may increase possibilities for learners to absorb the new language, while negative emotions may restrict potential language input.

However, studies on how mental well-being affects L2 learning are still in their infancy (see Dewaele, Chen, Padilla, & Lake, 2019; MacIntyre,Gregersen, & Mercer, 2019). Although positive self-concept and L2 self-efficacy (i.e., individuals' belief in their capacity to speak a second language) have been shown to be related for adults (Lake, 2016), the relationship between mental well-being and language performance has rarely been investigated for adult L2 learners. We hypothesize a difference in the association strength between mental well-being and language acquisition for learners in SA and AH contexts. Study-abroad learners are constantly confronted with challenges caused by insufficient language proficiency and cultural differences, obviously more so than at-home learners. These challenges may threaten their self-identities and trigger withdrawal due to high embarrassment potential (MacIntyre, 2002). Staying positive and resilient may be utterly important for international students. Therefore, it is conceivable that mental well-being may be more relevant to language learning for study-abroad learners than for at-home learners.

Language exposure

According to usage-based theories of language acquisition (Kemmer & Barlow, 2000; Bybee, 2013), the quality and quantity of language input or exposure an individual learner receives influence language learning outcomes. Studies have shown positive correlations between mastery of words and grammatical constructions and how often these words or constructions occur in language input (Tomasello, 2000; Bybee & Hopper, 2001; Ellis, 2002). However, it is unclear under what conditions and to what extent adult learners can learn from L2 exposure (see Carroll, 2001 for discussion). As the availability of L2 exposure depends on both learning context and on how much an individual seeks L2 engagement (Mitchell, Tracy-Ventura, & McManus, 2017), L2 exposure is bound to show great individual variability. Since SA and AH learning contexts differ in type, quality and quantity of L2 exposure, we are interested in investigating whether learning context mediates the association between amount of exposure and L2 learning.

It has proven quite a challenge to properly quantify L2 exposure for adult L2 learners (Collentine, 2009). Several questionnaires have been proposed to gauge L2 exposure, such as the language contact profile (Freed, Dewey, Segalowitz & Halter, 2004) and language engagement questionnaire (McManus, Mitchell, & Tracy-Ventura, 2014). These questionnaires have both been used to gather information on time spent reading, writing, listening, and speaking the L2 (e.g., watching movies, reading newspapers). The former asks for fined-grained estimates (e.g., minutes or hours) for language-related activities, whereas the latter asks participants to provide rough estimations of how frequently (e.g., rarely, several times a week) they are engaged in language-related activities. The present study employed the language engagement questionnaire.

Social interaction

SLA researchers (e.g., Long, 1996; Gass & Mackey, 2007) have argued for the importance of social interaction in language acquisition by stressing the benefits of negotiating meaning (e.g., by checking comprehension, repeating, rephrasing or requesting clarification). During the negotiation of meaning, interlocutors may have to make interactional modifications during conversation (e.g., slowing down speech rate, speaking more clearly, and simplifying sentence structures), which may not only help make input comprehensible but may also highlight linguistic features that might otherwise go unnoticed. Noticing a problem pushes learners to modify and improve their output (Schmid, 2007; Svalberg, 2007). In this way, interaction connects input, selective attention, and output (Gass, 2002). Previous studies have produced mixed outcomes on the relationship between interaction and language acquisition. Whereas some studies found that interaction and negotiation of meaning facilitated L2 vocabulary acquisition (de la Fuente, 2002; Long, 1996) and pragmatic competence (Taguchi, Xiao, & Li, 2016), others reported no significant relationship between interaction and oral fluency gains (e.g., Freed, Segalowitz, & Dewey, 2004).

Learning contexts may differ in opportunities for (spoken) social interaction in learners' L2. Immersion in an SA context allows learners to interact and negotiate in the target language. However, going abroad does not guarantee full immersion (Diao, Freed, & Smith, 2011). Degree of immersion is hence determined by the individual learner's socialization in the host community. Multiple factors affect the socialization process of study-abroad students, including their social capital, personality, language proficiency, and cultural differences between their country of origin and the host country (Kinginger, 2017; McManus, Mitchell, & Tracy-Ventura, 2014). Coleman (2013) found that international students' socialization process shows a general pattern. Socialization usually starts with students connecting to compatriots, then reaching out to other international students of different nationalities, and finally reaching the locals. As communication across linguistic and cultural boundaries is challenging, the socialization process may stagnate or even decline at any stage.

Measuring L2 interaction and socialization is challenging, similar to measuring L2 exposure as argued above. Attempts to quantify L2 interaction have largely been restricted to administration of questionnaires that ask participants to estimate the size and nature of their social network and time spent talking with people, such as the social network questionnaire (Mitchell, Tracy-Ventura, & McManus, 2017) and the Study Abroad Social Interaction Questionnaire (Dewey, Belnap, & Hillstrom, 2013). The present study adapted the Study Abroad Social Interaction Questionnaire whereby we focused on how long study-abroad learners had talked with whom in what language.

4.1.3 The current study

In order to observe how five individual-difference factors (i.e., language aptitude, working memory, mental well-being, language exposure, and social interaction) and learning context affect L2 listening proficiency, we invited participants from three learning contexts (i.e., SA context, AH-intensive context, and AH-regular context) to participate in a battery of L2 listening tests. These tests were administered at the beginning of their postgraduate program and again after one academic year. Participants also took individual-difference tests and questionnaires once at post-test. While SA context refers to studying abroad in a country where the target language is spoken as a native language, both the AH-intensive and AH-regular contexts refer to studying in the learner's country of birth where the targe language is learnt as a foreign language at school. Participants in the AH-intensive context learned the second language as a major, while those in the AH-regular context were majoring in other disciplines and only learned English as a subject. The tests and questionnaires in the study measure individual-difference factors described above and L2 listening proficiency. L2 listening proficiency is indicated by auditory vocabulary size and listening processing efficiency in the domain of lexical access, grammatical processing, and semantic processing. Processing efficiency is operationalized as a multi-dimensional construct comprising accuracy, speed, and stability of processing, which are measured by accuracy rate, reaction time, and coefficient of variation, respectively (see Segalowitz, 2010). We will use this set of measures (i.e., vocabulary size, reaction time, coefficient of variation, and accuracy rate) to reflect listening proficiency changes over the course of an academic year, and we will further observe how the proficiency measures at hand are associated with individual-difference factors in specific learning contexts.

Our research questions were the following:

1. Which individual-difference factors (i.e., aptitude, working memory, mental well-being, language exposure, and social interaction) are associated with listening proficiency and proficiency development across the three learning contexts, i.e., SA, AH-regular, and AH-intensive contexts?
2. Is there any difference between these learning contexts in terms of the strength or the direction in which proficiency is associated with these individual-difference factors?

4.2. Methods

4.2.1 Participants

143 Chinese postgraduates studying either in the UK (as newcomers) or China participated in this study. Participants were categorized into three groups: the SA group included 47 non-English-major students studying in the UK (M_{age} = 23.3, *SD* = 1.1), the AH-regular group included 51 non-English majors studying in China (M_{age} = 23.2, *SD* = 1.0), and the AH-intensive group included 45 English majors studying in China (M_{age} = 23.4, *SD* = 1.4). All participants were aged between 18-28 years, thereby minimizing the potential effect of age, an extra-linguistic factor, on time-sensitive measures and/or on learning in general. All participants had just finished bachelor education in China *before the pre-test*, with no previous study-abroad experience. The AH-regular and AH-intensive groups differed

mainly in terms of their majors. Majoring in English means that the classes of the AH-intensive group were mostly in English (around 1620 hours in total during their bachelor programs), while the AH-regular group majoring in non-English subjects had only one English class every week (around 144 hours in total during their bachelor programs). The SA group is a special group in that, though they were majoring in non-English subjects (around 144 hours of English instruction in total during their bachelor programs), they had to learn English in China in their spare time to prepare for study abroad in an English-speaking country.

Between our pre- and post-test, the AH-intensive group continued having English-medium courses (mostly taught by Chinese teachers; around 468 hours in total) but no basic language learning courses. The SA group also had English-medium courses (mostly taught by native English speakers; around 420 hours in total) and no basic language learning course. The AH-regular group had a two-hour college English class every week (around 72 hours in total). Therefore, though the SA group was not majoring in English, both the SA and the AH-intensive group would have English-medium education in the coming academic year. Contrastively, only the AH-regular group would mainly have Chinese-medium education. Additional information on participants' general language proficiency in terms of standardized test scores can be found in Appendix E.

4.2.2 Materials

Listening proficiency test

The listening proficiency test, administered at the pre- and post-test, consisted of several tasks: a lexical access task, a grammatical processing task, a semantic processing task, and an auditory vocabulary size test. The first three tasks (introduced below) were timed decision tasks, which we used to examine language processing efficiency on lexical, grammatical and semantic levels. Measures of language efficiency included accuracy rate, reaction time (RT) and coefficient of variation (CV), which were taken to index accuracy, speed, and stability of processing, respectively. Note that changes in speed and stability of processing have been argued to indicate different types of changes in learners' underlying processing mechanisms, with speed indicating quantitative changes (i.e., across-the-board speeding up) and stability indicating qualitative changes (i.e., a restructuring of processes) (Segalowitz, 2010). CV is a derived measure of RT and is calculated per individual as standard deviation of RTs divided by mean RT (Segalowitz, 2010). In addition, we employed an untimed multiple-choice task yielding a measure of declarative language knowledge, namely, auditory vocabulary knowledge.

1) Lexical access task

The lexical access task measured how well participants could recognize words by matching auditory and visual stimuli. On each trial, participants saw a picture and heard a word at the same time. They were asked to judge as fast as possible whether the picture and word matched or not, by pressing a corresponding button on a button box. If a picture and a word did not match, their referents shared phonetic similarities (e.g., "kite" and "cat"), fell into the same semantic category (e.g., "apple" and "orange"), or were unrelated (e.g., "frog" and "doctor"). This test contained six training trials and sixty experimental trials (half matching and half mismatching). Cronbach's alpha indices for the accuracy measures were .60 (pre-test) and .58 (post-test), while those for the RTs were .97 (pre-test) and .96 (post-test). The lower alphas for accuracy may relate to the (intended) relative ease of the tasks to make sure that participants would respond accurately to most items to ensure valid RT measurement for the speeded-response tasks.

2) Grammatical Processing Task

The grammatical processing task, including six training trials and sixty experimental trials, tapped into the processing of particular grammatical structures. Participants listened to a sentence and saw two pictures simultaneously, and were asked to quickly choose, by pressing a corresponding button, the picture that matched the sentence they had heard. To make the correct choice, participants had to understand the grammatical information of that sentence. For example, participants heard the sentence "the sheep eats" and saw two pictures, one with three sheep eating grass and the other with only one sheep eating grass. The 3rd person singular "-s" in this case is the grammatical cue leading to the correct choice. The grammatical cues of the sixty stimuli can be evenly put into three categories: morphological cues (plural "-s", 3rd person singular "-s", tense, and aspect), syntactic cues (dative, passive and cleft constructions, and relative clauses), and function words (prepositions, conjunctions, and pronouns). The twenty stimuli of the function-word category (out of the total of sixty) was later excluded from the analysis so that this task could better qualify as an indicator of grammatical/morpho-syntactic processing.

The stimuli were partly adapted from Kersten (2010), Waters, Caplan, & Rochon (1995) and Weist (2002). Cronbach's alpha indices for the accuracy measures were .57 (pre-test) and .55 (post-test), while those for the RTs were .92 (pre-test) and .93 (post-test). The alphas are lower for accuracy for reasons mentioned above (see Lexical access task).

3) Semantic Processing Task

The semantic processing task, taken from Lim and Godfroid (2015), measured how efficiently participants could form a semantic interpretation of a spoken sentence. Participants were asked to quickly indicate whether a spoken sentence was plausible or not. If a sentence was implausible, it violated either obvious factual knowledge (e.g. "A horse is an animal that can fly") or logic (e.g. "If you eat too much, you can get too thin"). This task consisted of six practice trials and fifty experimental trials (half of which were plausible and half implausible). Cronbach's alpha indices for the accuracy measures were .82 (pre-test) and .79 (post-test), while those for the RTs were .95 (pre-test) and .95 (post-test).

4) Vocabulary Size Test

The auditory vocabulary size test is a computerized version of the Peabody Picture Vocabulary Test Fourth Edition (PPVT™-4) (Dunn & Dunn, 2007). Participants heard a word and saw four pictures on a computer screen, and were asked to choose the picture that matched the word they had just heard. A total of 228 test items are grouped into 19 sets of 12 words, which are arranged in order of increasing difficulty. Test administration ended automatically if participants had made more than eight errors in one set. Unlike the first three tasks, the vocabulary size test is not timed. Participants could listen to a word multiple times if necessary. The PPVT has a reported average split-half reliability coefficient of .97 (Dunn & Dunn, 2007).

Individual-differences tests and questionnaires

The individual difference tests included three LLAMA language learning aptitude subtests, a working memory test, and questionnaires collecting various types of information (i.e., language exposure, social interaction, mental well-being, and background information).

1) LLAMA language learning aptitude tests

a. LLAMA_B associative learning test

The LLAMA_B test (Meara, 2005) measures associative vocabulary learning ability. Participants were given 120 seconds to memorize twenty pictures and their corresponding names. The name of a picture would only show with a click on the picture, and would disappear immediately after the participant clicked on another picture. The set of twenty pictures were always on screen. Participants would hear a beep sound when the 120 seconds were done. Afterward, participants were required to choose the corresponding picture according to the name prompt they saw one at a time. The twenty name prompts would appear on the screen one by one in random order, and would only disappear after participants had made a choice by clicking on a picture. Depending on the number of correct answers, the score for the LLAMA_B test could be between 0 and 100 (i.e., 5 points allocated for each item).

b. LLAMA_D phonetic decoding test

The LLAMA_D test (Meara, 2005) measures phonetic decoding ability. Participants were required to listen carefully to audio clips of ten unfamiliar words (stimuli based on the native languages of North West British Columbia). Each word could only be played once. The time interval between two adjacent words was two seconds. A beep sound signified the end of the learning phase. Then participants would listen to another twenty words one by one and indicate which of these words had been heard previously. Depending on the number of correct answers, the score for the LLAMA_D test could be between 0 and 100 (i.e., 5 points allocated for each item).

c. LLAMA_F grammar inference test

The LLAMA_F test (Meara, 2005) measures grammar inference ability. Participants were given five minutes to study the grammar of an unknown language. They could access twenty pairs of pictures and sentences, by clicking twenty corresponding buttons on the computer screen. Within the given time, participants could click any of these buttons to compare and contrast the pictures and sentences to work out the underlying grammatical rules. After the end of this learning phase, they would hear a beep sound. Afterwards, participants were presented with a new set of 20 pictures, each of which was accompanied by two sentences. They were asked to choose which sentence matched the picture according to the grammatical rules they had just learned. Depending on the number of correct answers, the score for the LLAMA_F test could be between 0 and 100 (i.e., 5 points allocated for each item).

2) Working memory test

To assess auditory working memory capacity, a digit span test (Soylu, 2010) was employed, consisting of a section with forward digit sequence recall and one with backward recall. In the forward section, participants heard an increasingly longer sequence of digits (i.e., from 0 to 9 in their native language) and then were asked to type these digits on the keyboard in the same order as presented. They would hear two different series of digits of the same length. If they could not repeat correctly either time, this section would be automatically stopped. Otherwise, they would continue to hear new digit sequences with one more digit in length. The number of correct answers given was their score in this section. The backward section was the same as the forward section except that participants were to repeat the digits in the reverse order. This task yields two scores (one for forward recall and another for backward recall) based on the number of correct answers given in either section.

3) Language engagement questionnaire

The language engagement questionnaire was adapted from McManus, Mitchell, and Tracy-Ventura (2014). It contains 22 items, each of which asked how frequently one participated in a certain activity in English (e.g., listening to music, writing emails, browsing the internet, or having a longer than 5-min conversation). Answers had to be given on a five-point scale (i.e., never, rarely, a few times a week, several times a week, every day), with higher numbers indicating more English language engagement.

4) Social interaction questionnaire

The social interaction questionnaire, adapted from Dewey, Belnap, and Hillstrom's (2013) Study Abroad Social Interaction Questionnaire, was used to document participants' social network and interaction. Participants were required to fill in whom they had talked with in the past two weeks, the language status of their interlocutors (i.e., native Chinese speaker, native English speaker, other English-speaking international), in which language the conversations were conducted, and how many minutes they had talked with each interlocutor. The total number of minutes of their contribution to the conversations (irrespective of whether they talked with a native or nonnative speaker) therefore indicate their degree of social interaction and integration.

5) Mental well-being questionnaire

The short-version Warwick-Edinburgh Mental Well-being Scale (WEMWBS) was used to measure different aspects of mental health, yielding an indicator of positive emotions our participants were experiencing. It consists of five statements about positive affect, satisfying interpersonal relationships and positive functioning (Tennant et al., 2007). Participants were required to evaluate the extent to which each statement matches their own experience on a five-point scale (i.e., never, rarely, sometimes, often, always). The higher score they got, the more positive they felt.

6) Background questionnaire

Apart from the previously described questionnaires, participants were also asked to fill in background information such as age, gender, language learning experience (i.e., years of English education), and standardized English proficiency test scores (see section 4.2.1 above for more information).

4.2.3 Procedures

Data collection

The first author recruited participants at universities in mainland China and London area. Participants had to first fill in a screening questionnaire. Eligible participants (see section 4.2.1 for recruitment criteria) were invited to take a series of listening tests twice with an interval of seven months (roughly corresponding to the duration of an academic year), as well as to take the individual-difference tests and questionnaires once at post-test. In the first round of data collection, the researcher explained that this study had been approved by the Ethics Assessment Committee of Institution X and that the collected data would only be used for research purposes. After a detailed description of all the tasks, the researcher asked participants to sign an informed consent form before the administration of the test. They received a small monetary compensation for their participation at both pre-test and post-test.

Data cleaning criteria

Firstly, items in the language processing tasks were excluded as ambiguous items if they elicited accuracy rates below 80% in a native-speaker sample (Chapter 2). Secondly, participants were excluded when their accuracy rate was below 50% (chance level) on any speeded test, or their vocabulary score was not within three standard deviations of the group mean, or when they failed to do all tasks. Thirdly, only RTs of correct responses were analyzed. Fourthly, RTs below 250 milliseconds (measured from audio onset) were removed as invalid responses.

Statistical analysis

To reduce the number of individual-difference factors in the main analysis, we first conducted a confirmatory factor analysis and estimated the latent variables of the several individual differences measures. These latent variables were to be fitted into mixed models as fixed-effect predictors, which allowed us to look into the effect of individual-difference factors on language learning and whether these factors interacted with learning contexts.

4.3 Results

The research questions will be addressed by a series of mixed-effects regression models, which have task performance as dependent variable and Task (i.e., grammatical processing vs. lexical access task; semantic processing vs. lexical access task), Group (i.e., AH-regular vs. SA group; AH-intensive vs. SA group), Time (i.e., pre- vs. post-test), and individual-indifference variables as independent variables. Note that the vocabulary size task is not included in the Task variable because this task is analysed separately from the three speeded-response processing tasks. Before introducing the results of our main analyses, the descriptive statistics of task performance and observed individual-difference factors will be presented, and the structure of individual-difference factors will be investigated by a confirmatory factor analysis (CFA).

Table 4.1 displays the descriptive statistics of the task performances of the participant group in the pre- and post-test. Table 4.2 shows the descriptive statistics of individual-difference factors collected with cognitive tests and self-reported questionnaires.

Table 4.1: Descriptive statistics of task performance in the pre- and post-tests.

Task	Measure	Pre-test	Post-test
		Mean (SD)	*Mean (SD)*
Lexical access task	Accuracy	.86 (.06)	.86 (.05)
	RT	1207 (276)	1034 (205)
	CV	.34(.09)	.31 (09)
Grammatical processing task	Accuracy	.85 (.07)	.87 (.07)
	RT	3307 (440)	2941 (410)
	CV	.38 (.05)	.39 (.05)
Semantic processing task	Accuracy	.90 (.10)	.92 (.08)
	RT	3236 (580)	2837 (481)
	CV	.28 (.06)	.25 (.06)
Vocabulary size test	Vocabulary size score	131 (26)	136 (25)

Note: The values of Accuracy are proportions and RT values are in milliseconds.

Table 4.2: Descriptive statistics of the eight observed individual-difference variables

Measure (scale or unit)	*Mean*	*SD*
LLAMA_B associative learning test (0-100)	48.64	19.43
LLAMA_D phonetic decoding test (0-100)	35.63	13.13
LLAMA_F grammar inference test (0-100)	62.80	20.84
Forward digit span (number of digits)	14.50	3.34
Backward digit span (number of digits)	14.17	3.45
Language engagement (1-5)	2.38	0.74
Mental well-being (1-5)	3.65	0.61
Speaking time (log-transformed minutes)	3.11	2.27

Note: As the original Speaking time variable was heavily skewed (the mean and standard deviation were 134 and 280), it was log-transformed to normalize its distribution before it was fitted into the confirmatory factor analysis model.

4.3.1 Confirmatory factor analysis: Structure of the individual-difference variables

To limit the number of individual-difference factors, LLAMA_B, LLAMA_D, and LLAMA_F were reduced to an Aptitude factor. Forward digit span and Backward digit span were reduced to a Working memory factor. This led to a five-factor model for the individual-difference variables (see Figure 4.1), which was tested in a confirmatory factor analysis (CFA). The fit indices of the model ($\chi2$ (13) = 9.369; *CFI* = 1.000; *TLI* = 1.074; *SRMR* = 0.036; *RMSEA* = 0.000) indicates that this model is a good representation of the individual-differences data. An attempt to further cluster Language engagement and Speaking time into one latent variable was not successful. Language engagement, Speaking time, and Mental well-being were therefore kept separate, corresponding to the factors Exposure, Social interaction, and Emotion, respectively, in the CFA model. Intercorrelations among observed variables are reported in Appendix Table F1.

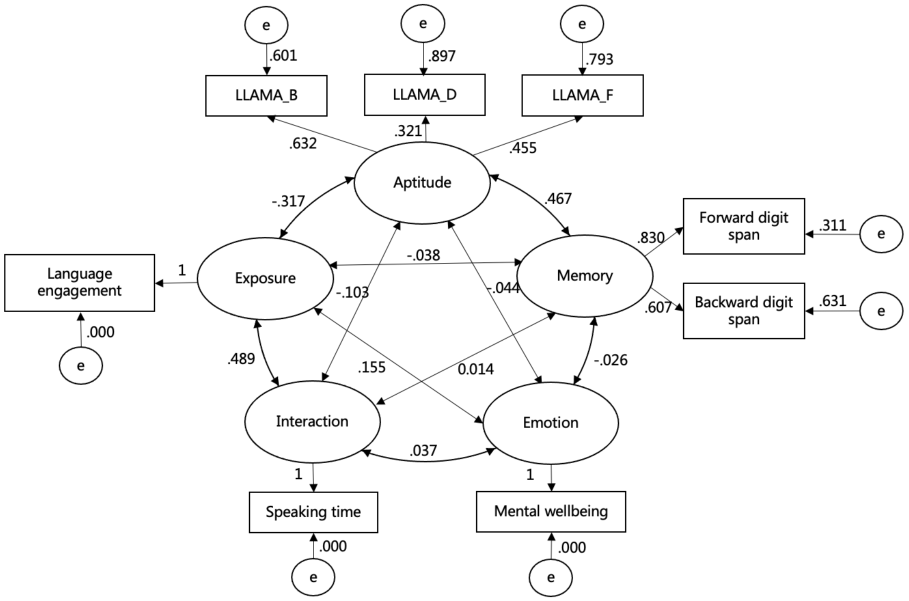

Figure 4.1: Standardized factor loadings, correlations, and error variances of the individual-difference factors in the CFA model.

4.3.2 Main analyses: Relationship between language proficiency, learning contexts and individual-difference factors

Mixed-effects regression models[1] were fitted to predict ACCs, RTs, CVs, and vocabulary sizes. We started fitting models with maximal fixed-effects structures, by assuming the existence of all the interactions between fixed-effect predictors (i.e., Time, Group, Task, and Individual-difference factors), as well as with two random intercepts (i.e., for Subject and Item). Higher-order interactions may, however, not have a significant effect on the dependent variables. Therefore, we then compared the maximal fixed-effects models with models from which all four-way, three-way and/or two-way interactions were successively deleted in order to identify the optimal models, thereby at the same time reducing the risk of capitalizing on chance (see Appendix Table F2). After we decided on the optimal models through chi-square difference tests, we maximized their random-effects structure (i.e., all possible random by-Subject and by-Item slopes for the fixed effects Group, Task, and Time) on the premise of model convergence, in order to further improve model fits. The specifications of the four regression models can be found in the appendix from Table F3 to Table F6, respectively.

Subsequent sections describe the model results. The main effects of individual-difference factors, the interactions between Time and individual-difference factors, the interactions between Time, Group, and individual-difference factors, and the interactions between Time, Group, Task, and individual-difference factors are most critical for answering the research questions of this study and will be described in detail.

Vocabulary size

A chi-square difference test shows that a main-effect model fitted the vocabulary-size data best, indicating no significant interactions between Time, Group, and individual-difference factors in the vocabulary size model (see Appendix Tables F2 and F3). Participants' performance in the vocabulary size test improved significantly during the academic year, as evident from a significant effect of the Time variable ($\beta = 4.63$, $SE = 1.16$, $t = 4.01$, $p < .001$). The AH-intensive group outperformed the SA group ($\beta = 13.22$, $SE = 5.64$, $t = 2.34$, $p = .021$), who in turn outperformed the AH-regular group ($\beta = -19.72$, $SE = 6.15$, $t = -3.20$, $p = .002$) in the pre-test. There was no significant difference among these three groups in terms of vocabulary gains over the academic year, as there was no interaction between Time and Group effects.

With regard to individual-difference factors, Aptitude manifested a positive effect on vocabulary knowledge in all the learning contexts ($\beta = 9.84$, $SE = 2.75$, $t = 3.57$, $p < .001$), without significant interaction between Aptitude and learning context. Effects of WM, Exposure, Social interaction and Emotion on vocabulary size failed to reach significance. Individual-difference factors did not predict improvement made by participants over the academic year, nor did it account for any potential between-group differences in vocabulary size, as there were no interactions between these variables and the Time or Group effects.

Processing accuracy

A chi-square difference test showed that a two-way interaction model fitted the accuracy data best, indicating no significant three-way or four-way interactions between Time, Group, Task, and individual-difference factors. Participants in general improved in processing accuracy at post-test compared to pre-test (β = 0.20, *SE* = 0.04, *z* = 4.56, *p* < .001) (see Appendix Tables F2 and F4). The SA group did not differ significantly from the AH-intensive group (β = 0.21, *SE* = 0.18, *z* = 1.16, *p* = .248), but significantly outperformed the AH-regular group (β = -0.62, *SE* = 0.20, *z* = -3.04, *p* = .002) in processing accuracy at pre-test. The three groups did not differ from each other significantly in terms of their improvement in processing accuracy over the academic year, as indicated by non-significant interactions of Group by Time (β = -0.22, *SE* = 0.12, *z* = -1.91, *p* = .056 ; β = -0.04, *SE* = 0.11, *z* = -0.35, *p* = .730).

With regard to individual-difference factors, Aptitude (β = 0.36, *SE* = 0.07, *z* = 4.87, *p* < .001) and Exposure (β = 0.07, *SE* = 0.03, *z* = 1.98, *p* = .048) had a positive effect on processing accuracy across the tasks, but the effects of WM, Social interaction and Emotion failed to reach significance. Similar to the vocabulary model, the processing accuracy model did not show significant interactions between these variables and the Time or Group effects, indicating that individual-difference factors did not correlate with improvement made by participant over the academic year nor explain the between-group differences in processing accuracy.

Processing speed

A chi-square difference test showed that a four-way interaction model fitted the reaction time data best, indicating complex interactions between Time, Group, Task, and individual-difference factors (see Appendix Table F2). The model output showed that several four-way interactions were significant (between Time, Group, Task and one of the individual differences variables). To follow up on these four-way interactions, we analysed the results per task to check for which task we observed interactions between Time, Group and any of the individual differences variables. These per-task analyses showed that none of the three-way interactions between Time, Group and individual differences variables reached significance (see Appendix Table F5). This implies that tendencies may exist for the amount of improvement over time, as related to individual-difference factors, to differ between groups and tasks, but these tendencies are not strong enough to show up for any of the three tasks.

Participants generally responded faster at post-test than at pre-test for the lexical access task (*β* = -0.14, *SE* = 0.03, *t* = -5.05, *p* < .001). The same also held for the grammatical processing task (*β* = -0.13, *SE* = 0.02, *t* = -6.41, *p* < .001) and the semantic processing task (*β* = 0.14, *SE* = 0.02, *t* = -7.50, *p* < .001). The only significant between-group difference in RTs at pre-test was in the lexical access task where the SA group was significantly faster than the AH-regular group (*β* = 0.21, *SE* = 0.08, *t* = 2.72, *p* = .008) but did not differ from the AH-intensive group (*β* = 0.05, *SE* = 0.06, *t* = 0.79, *p* = .432). The participant groups did not differ in terms of their improvement in processing speed during the academic year except that the AH-regular group made significantly more progress than the SA group in the lexical access task (*β* = -0.18, *SE* = 0.07, *t* = -2.49, *p* = .014).

With regard to individual-difference factors, aptitude had a positive effect on processing speed (i.e., shorter RTs) in the grammatical processing (*β* = -0.04, *SE* = 0.02, *t* = -2.18, *p* = .031) and the semantic processing tasks (*β* = -0.05, *SE* = 0.02, *t* = -2.52, *p* = .013), but not for the lexical access task. Similarly, the amount of language exposure predicted processing speed in the grammatical processing (*β* = -0.02, *SE* = 0.01, *t* = -2.36, *p* = .020) and the semantic processing tasks (*β* = -0.02, *SE* = 0.01, *t* = -2.37, *p* = .019), but not for the lexical access task. The effects of Working memory, Social Interaction, and Emotion failed to reach significance for any of the speeded-response tasks. Meanwhile, there were no significant two-way or three-way interactions between Time, Group, and individual-difference factors in any of the per-task models, meaning that the effect of individual-difference factors on improving processing speed did not differ across groups or tasks.

Processing stability

A chi-square difference test showed that a two-way interaction model fitted the CV data best, indicating no significant three-way or four-way interactions between Time, Group, Task, and individual-difference factors (see Appendix Tables F2 and F6). Participants' RTs at post-test, in general, were more stable than those at pre-test (*β* = -0.02, *SE* = 0.00, *t* = -5.31, *p* < .001). RTs in the lexical access tasks were more stable than those in the grammatical processing task (*β* = 0.06, *SE* = 0.01, *t* = 9.77, *p* < .001), but less stable than those in the semantic processing task (*β* = -0.07, *SE* = 0.01, *t* = -11.73, *p* < .001). The progress participants made in the lexical access task was larger than that in the grammatical processing task (*β* = 0.04, *SE* = 0.01, *t* = 4.91, *p* < .001), but did not differ from that in the semantic processing task (*β* = 0.01, *SE* = 0.01, *t* = 1.07, *p* = .284). However, no effect of the individual-difference factors was found on CV.

4.4 Discussion

Our first research question addressed the effects of individual-difference factors on L2 learning. We found that aptitude was the only individual-difference factor that predicted vocabulary size scores, and it did so across learning contexts. Both aptitude and language exposure correlated with processing efficiency, more specifically, with processing accuracy across the three speeded-response tasks, and with processing speed in the grammatical processing and the semantical processing tasks. Working memory, mental well-being, and social interaction did not relate to any of the listening measures. As for improvement made by participants on the listening measures over the academic year, no effects of individual-difference factors were observed in any of the learning contexts. Note that, although processing stability (as measured by CV) was found to be sensitive to the effects of Time, Task and Group variables, none of the individual-difference factors was associated with processing stability.

Our second research question examined the potential interplay between individual-difference factors and learning context. We did not find evidence that learning context mediated the way individual-difference factors related to listening proficiency or its progress. The rest of this section will break down these results, reviewing and discussing the effect of each individual-difference factor on second language learning as well as their potential interaction with learning context in detail.

4.4.1. The effects of language aptitude and working memory on second language learning

We found that aptitude was predictive of participants' vocabulary size and processing efficiency (i.e., of processing accuracy in all three speeded-response tasks, and of processing speed in the grammatical processing and the semantic processing tasks) across learning contexts. The association between aptitude and vocabulary size among our participants makes sense in light of the relationship between associative learning ability, as a subcomponent of language aptitude, and as measured by the LLAMA_B test, and vocabulary acquisition. Associative learning ability indicates how well learners can link communicative signals (e.g., words, sound, and pictures). Previous studies have indeed confirmed that aptitude predicts initial level of success for vocabulary learning in a novel language. For example, Dahlen and Caldwell–Harris (2013) asked English speakers to memorize twenty Turkish words and their corresponding pictures (with 20 seconds of rehearsal for each word), and found that participants with higher aptitude could recall and recognize more words than those with lower aptitude. The present study agrees with these results and further provides evidence that aptitude also predicts vocabulary size for intermediate to advanced learners.

Moreover, our results expand on earlier studies on the relationship between aptitude and language processing (e.g., Suzuki & DeKeyser, 2017; Yi, 2018) by observing a positive relationship between aptitude and processing efficiency in the domains of lexical access, grammatical processing, and semantic processing. Yi (2018) found that language aptitude measures (i.e., LLAMA_B and LLAMA_F) predicted processing accuracy of collocations, measured by a timed phrasal acceptability judgment task. Yi accounted for these results by suggesting that aptitude may relate to semantic processing of larger-than-word units. Suzuki and DeKeyser (2017) reported that grammar inferential ability, as measured by the LLAMA_F test, predicted performance in timed grammaticality judgment tasks among advanced Chinese learners of Japanese. Their study provided evidence that aptitude in grammar inference related to grammatical processing for advanced L2 learners. The present study, given its latent-variable approach, was not set up to pinpoint which components of aptitude correspond exactly to which aspects of language processing. However, our results support a stable relationship between aptitude and language processing, suggesting that aptitude does not only relate to crystallized measures of language proficiency, but also to the efficiency with which the L2 is processed.

Furthermore, our vocabulary and language processing results did not show any significant interactions between Time and Aptitude. That is, aptitude predicted overall listening proficiency at pre- and post-test, but did not predict our participants' language progress over the course of an academic year. This, nevertheless, seems to suggest that aptitude is still a stable predictor of language proficiency for intermediate to advanced learners. Therefore, the present study did not provide evidence to support Robinson (2013)'s view that "with growing L2 proficiency, the learner relies less on the basic cognitive skills associated with FL aptitude for continued progress" (see also Winke, 2013).

Surprisingly, we did not observe any effect of working memory on any of the listening proficiency measures, whereas current literature suggest that L2 acquisition and processing seem to be intertwined with working memory (for a meta-analysis, see Linck et al., 2014; for a narrative review, see Williams, 2012). Higher memory capacity might open a larger processing window for language sequences (Janacsek & Nemeth, 2013). However, working memory and aptitude, both used as predictors in the present study, may have a considerable amount of shared predicted variance (see Mackey, Adams, Stafford, & Winke, 2010). Therefore, the lack of correlation between working memory and L2 performance in our study is likely due to the co-existence of aptitude and working memory in our statistical models[2].

4.4.2. The effects of exposure and interaction on second language learning

We found that language exposure was associated with processing efficiency. Exposure was predictive of processing *accuracy* in all three tasks and exposure predicted processing *speed* in the grammatical processing and the semantic processing tasks. This suggests that the more language exposure learners have, as quantified by the language engagement questionnaire, the more efficient they are at language processing. This association can be explained by the central concept of the skill acquisition theory of second language acquisition, stating that reaction time and error rate decrease with practice (DeKeyser, 2015). Social interaction, on the other hand, was not significantly related to any of the listening measures, suggesting that L2 learners with higher amount of social interaction do not necessarily have higher listening proficiency. Hence this study does not provide evidence that social interaction facilitates L2 listening development. This result may be explained, to some extent, by DeKeyser's (2015) skill acquisition theory of second language acquisition, which claimed that language learning practice is skill-specific. Social interaction, as a measure in our study, is language production in interaction. Previous research has also provided empirical evidence that input-based practice is more beneficial for receptive skills (i.e., listening and reading) and may have limited transfer to the improvement of productive skills (i.e., speaking, and writing), and vice versa (De Jong, 2005; DeKeyser & Sokalski, 2001; Rodgers, 2011). Input-based learning practice (e.g., listening to broadcast, watching TV shows, and attending to lectures) may be more effective means for advancing their listening skills than social interaction. Furthermore, the questionnaire of social interaction required participants to estimate their speaking time with others in minutes. Such fine-grained time estimations may not be easy for participants to conceptualize and thus may be unreliable. Hence the null effect of social interaction ought to be interpreted with caution.

4.4.3. Mental well-being and second language learning

We found no relationship between mental well-being (i.e., Emotion) and second language proficiency, and this observation held across learning contexts. Possibly, the reason why our study did not provide any evidence to support our initial hypothesis about the relationship between mental well-being and second language acquisition was that we approached participants while they were preparing end-of-semester exams. The extra pressure from preparing exams might have, to some extent, affected their mental well-being, or might have obscured any differences among participants. Relatedly, the questionnaire we used might not have been sensitive enough to detect differences in mental well-being among the participants, who might have been unwilling to disclose their personal feelings. Meanwhile, questions as to whether and how mental well-being relates to second language proficiency are rarely investigated. Future studies are encouraged to further investigate the relationship between mental well-being and second language acquisition in other contexts.

4.4.4 The interaction between individual differences and learning context

As for any interactions between individual-difference factors and learning context, we found none. Previous studies on the interaction between individual differences and learning context are scarce and have reported mixed results (e.g., Sunderman & Kroll, 2009; Tokowicz et al., 2004; Faretta-Stutenberg & Morgan-Short, 2018). For example, some studies (e.g., Sunderman and Kroll, 2009; Tokowicz et al., 2004) have reported a positive relationship between working memory and L2 in SA contexts, but not for learners in AH contexts. These authors argued that WM should play a larger role in SA contexts where processing demands are arguably higher (McDonald, 2006). Conversely, some other studies found that working memory predicted language learning in explicit learning conditions, such as regular classroom setting in AH contexts (e.g. Linck and Weiss, 2015; Sagarra, 2008). They argued that higher working memory give learners a better chance to successfully retain metalinguistic information in memory while simultaneously comprehending and producing language (e.g. Linck and Weiss, 2015; Sagarra, 2008). It should be noted, however, the studies that have reported an interplay between individual differences and learning context only compared the role of individual differences in different learning contexts/conditions descriptively, instead of testing for statistical interactions between individual difference variables and learning context (e.g., Sunderman and Kroll, 2009; Tokowicz et al., 2004; Williams, 2012).

The present study found no significant interactions between individual differences and learning context, which is in line with results by for instance Robinson (1997) who also reported no interactions between individual-difference factors and (implicit or explicit) learning conditions. Robinson argued that adult second language learning in implicit and explicit learning conditions is fundamentally similar. If our findings hold, in that there is no qualitative difference in the way predictors of L2 learning relate to L2 outcomes in different learning contexts, our results provide supporting evidence for fundamental similarity of adult second language learning across learning contexts.

4.4.5 Limitations and strengths

One limitation of the present study is that the interval between pre- and post-test is only seven months, which roughly corresponds to an academic year. Studies covering longer periods between pre-test and post-test may allow better investigation of the relationships between individual differences, learning context and second language development. Future studies are encouraged to extend the study period allowing learners to show more language progress. Another limitation is that we administered the questionnaires indexing learners' mental well-being, language engagement, and their social interaction only at post-test. This single administration does not rule out the possibility that the information collected could be affected by (unknown) third factors at the time of testing (e.g., affected by exam stress or upcoming vacations), and may thus not be completely representative of the whole study period. Future studies of this kind are recommended to collect such information at multiple points in time.

As for its strengths, this study investigated how individual-difference factors and learning context jointly affect second language learning, allowing us to investigate possible interactions between individual-difference factors and learning context. Moreover, we have examined the effect of individual differences on L2 listening in a relatively extensive manner by investigating acquisition of language knowledge as well as processing efficiency, which has rarely been done in previous research. Future studies might want to extend this type of language-research approach to other types of populations or learning contexts.

4.5 Conclusion

This study has demonstrated how individual-difference factors (i.e., aptitude, exposure, and social interaction) relate to different aspects of L2 listening comprehension in SA and AH learning contexts. None of these individual-difference factors predicted improvement participants made over this academic year, but some were predictors of listening proficiency at both pre- and post-test. Language aptitude was the only individual-difference factor that predicted vocabulary size, and it did so across learning contexts. Concerning processing efficiency outcomes, aptitude and exposure predicted processing accuracy across tasks and predicted processing speed in two out of three tasks. The effect of working memory was not shown for any outcome measure, which may have been due to its relationship with aptitude (being the stronger predictor). Mental well-being and social interaction were not found to be predictive of any of the measures, either.

Importantly, no interactions between learning context and individual-difference factors were observed, which suggests that second language learning for adult L2 learners is fundamentally similar across learning contexts.

Notes

1. The ACC model was a logistic regression model (accuracy being a binary variable, coded as 0 or 1 for each individual response) and the other models (with continuous variables) were linear regression model. RTs were log-transformed before entering into the RT model to normalize their distribution. As CV was an aggregated measure calculated on task level for each participant, the CV models did not have any item-level variables.

2. To follow up on the intercorrelation between aptitude and working memory, we reran all the statistical models without aptitude. In the vocabulary size, processing accuracy, and processing speed models, working memory functioned in a similar way as aptitude did in the currently reported models regarding effect directionality and significance. This supports our speculation that the supposed relationship between language measures and working memory was taken away by that of aptitude in the currently reported models.

Chapter 5: The impact of studying abroad on L2 development: A multilevel meta-analysis

Abstract

To evaluate the impact of studying abroad on language learning, we reviewed twenty studies comparing L2 development in study-abroad and at-home learning contexts. A multi-level meta-analysis of 105 individual effects yielded a small-to-medium overall effect (favoring studying-abroad over at-home groups; g = 0.31). As all the studies included in our meta-analysis were of between-group pretest-posttest designs, we simulated the types of designs we did not include, to determine whether different choices of methodological designs would influence the magnitude of observed effect sizes. Results showed that between-group pretest-posttest designs provided the most conservative estimation of study-abroad effects, followed by within-group pretest-posttest designs, followed, in turn, by between-group posttest-only designs. Furthermore, moderator analyses indicated that long-term study-abroad programs were more effective in facilitating L2 development than short-term ones; and that, compared to their at-home counterparts, study-abroad learners showed more progress on general proficiency and processing-related measures than on knowledge measures. Although the overall study-abroad effect was significant for production and not for comprehension, the moderation effect of language dimension (comprehension vs. production) just missed significance. No moderation effects of language modality (auditory vs. visual) or cultural distance were observed.

5.1 Introduction

The popular belief that second language acquisition benefits more from a study-abroad (SA) learning context than from an at-home (AH) learning context is one of the driving factors for the increasing popularity of study-abroad programs (Freed, 1998; Tullock & Ortega, 2017). There is widespread interest among researchers in examining the effect of studying abroad on second language learning (e.g., Collentine 2004; Segalowitz & Freed, 2004; Serrano, Llanes, & Tragant, 2011; Sasaki, 2007; Llanes & Muñoz, 2013). A body of SA research has gradually formed over the past two decades, covering various aspects of second language learning (e.g., general language proficiency, vocabulary size, grammatical knowledge, oral fluency, writing fluency, pragmatic competence). However, instead of testing the comparative linguistic gains of studying abroad against staying in the home country over a period of time, SA studies, more often than not, only compared a group of learners before and after an SA experience (Serrano, 2010; Rees & Klapper, 2008). This lack of a comparison group means that the progress demonstrated in those studies might sometimes be ascribed to other factors (e.g., test repetition effects common to pretest-posttest designs and general improvement of proficiency over time irrespective of learning context), rather than the actual language-learning benefits of an SA context. Only a small number of studies among the myriad of study-abroad research involved a pairwise comparison of study-abroad and at-home groups with a pre- and post-test. Furthermore, findings on the relationship between studying abroad and second language proficiency are often inconsistent and unclear, which may, to some extent, be associated with variability in study characteristics (e.g., length of stay and type of measurement). That is, these factors may potentially moderate the direction and magnitude of observed study-abroad effects, accounting for some of the inconsistency in research findings.

To gain a deeper and more comprehensive understanding of the differential effects of SA and AH learning contexts on L2 development, this study performed a multilevel meta-analysis on previous SA studies of a between-group pretest-posttest design. We aim to quantitatively synthesize SA effects reported by these studies and to address the differences in research findings by exploring potential moderators of reported SA effects (e.g., length of study and specific aspects of language learning in question). Although studies of between-group pretest-posttest designs were meta-analyzed before by Yang (2016), the meta-analytical methods Yang employed only focused on the posttest rather than both the pre- and post-tests, thus reducing the original between-group pretest-posttest design to an equivalent of a between-group posttest-only design. This present study proposes to conduct a multi-level meta-analysis with a different approach, which takes into account between-group differences at pretest. It allows for the inclusion of effect sizes on all measures of the included studies, and hence also allows for more informative moderator analyses.

5.1.1 Overview of study-abroad research

Despite of the supposed superiority of an SA context over an AH context (e.g., in terms of quantity and quality of language exposure), the advantage of studying abroad in language learning does not clearly show from an overview of study-abroad research. Instead, research on study-abroad language learning revealed mixed and sometimes contradictory results. Some studies found that SA learners showed greater gains than AH learners in oral fluency and proficiency (Segalowitz & Freed, 2004; Freed, Segalowitz, & Dewey, 2004; Llanes & Muñoz, 2013; Serrano, Llanes, & Tragant, 2011; Jochum, 2014), narrative abilities and semantic density in oral production (Collentine, 2004), grammatical processing (Faretta-Stutenberg & Morgan-Short, 2018), nativelike accent (Muñoz & Llanes, 2014), pragmatic competence (Matsumura, 2001; Félix-Brasdefer, Hasler-Barker, 2015), writing fluency and proficiency (Sasaki, 2007, 2011; Serrano, Llanes, & Tragant, 2011), communicative strategies (Lafford, 2004; Montero, Serrano, & Llanes, 2017), and spoken-language processing efficiency (Chapter 3). Marginal or no differences between SA and AH learners were sometimes reported in grammatical knowledge (Collentine, 2004; Isabelli-Garcia, 2010; Håkansson & Norrby, 2010), oral and written fluency, lexical and syntactic complexity, and accuracy (Serrano, Llanes, & Tragant, 2011; Llanes & Muñoz, 2013) and receptive vocabulary size (Chapter 3). Finally, some studies even reported that AH learners made more progress than SA learners in discrete grammatical and lexical features in oral production (Collentine, 2004), written syntactic complexity (Llanes & Muñoz, 2013), adjective ordering (Hirakawa, Shibuya, Endo, 2019), and TOEFL paper-based test scores (Cutrone & Datzman, 2015). A summary of the research designs and results of these empirical studies can be found in Appendix Table G1 (for extensive narrative reviews on SA research, see e.g., DeKeyser, 2014; Kinginger, 2009; Llanes, 2011; Serrano, 2010; Pérez-Vidal, 2017).

5.1.2 Potential moderators of SA effects

The inconsistency in research findings might, to some extent, be ascribed to differences in study characteristics, such as methodological design, specific language skills in question, length of stay, type of measurement, cultural distance between the native and the host country, and participants' age and pre-departure L2 proficiency level (e.g., novice, intermediate, and advanced). These factors might moderate SA effects on second language learning, which will be elaborated on below. Note that, as the studies included in the present meta-analysis focused primarily on university students who show little variance in age (i.e., clustering around twenty years of age) and pre-departure proficiency level (i.e., mostly intermediate), the present study will not investigate the possible moderation effect of these two factors.

A. Methodological design

The methodological designs of quantitative SA research can be classified into three categories: between-group posttest-only designs, within-group pretest-posttest designs, and between-group pretest-posttest designs. A between-group posttest-only design compares SA learners with AH learners at a single point in time, assuming similar starting levels prior to the onset of a study period. Note that since SA learners probably learned the target language more intensively than their at-home counterparts to prepare for studying abroad, considerable between-group differences may already exist at pretest (see Chapter 3). Contrastively, studies adopting a within-group pretest-posttest design compare L2 proficiency of the same participants before and after an SA experience. As mentioned before, without an AH comparison group, this category of studies risks inflating the effect sizes of studying abroad, as they cannot distinguish general improvement of proficiency irrespective of learning contexts from the actual linguistic benefits of studying abroad (see also Serrano, 2010). Therefore, the last category of SA research, the between-group pretest-posttest design, is the most comprehensive and methodologically rigorous in that it compares both pretest and posttest performance of SA learners with that of AH learners. Note that this type of design is not flawless either, as it may suffer from non-randomized group assignment. That is, SA and AH groups may have pre-existing differences in many aspects (e.g., language proficiency, language learning motivation, and language aptitude), although these differences can be partially addressed by taking into account between-group differences at pre-test. In sum, it is likely that methodological design modulates the observed magnitude of SA effects across studies.

B. Specific linguistic subskills

According to DeKeyser's (2015) skill-based account of second language acquisition, language skills (listening, speaking, reading and writing) may not develop simultaneously or evenly in learners due to the skill-specific effect of language learning practice. That is, practice in one skill (e.g., reading) contributes to the improvement of this specific skill but not necessarily to the improvement of other seemingly related skills (e.g., writing, speaking, and listening). Skill specificity, as an important concept of skill acquisition theories (Anderson, 1983; DeKeyser, 2015), can be explained by the distinction between declarative knowledge, i.e., explicit knowledge stored in one's memory (e.g., vocabulary), and procedural knowledge, i.e., implicit knowledge assessed unconsciously in skill performance (DeKeyser, 2007). Procedural knowledge for each skill is highly specific and may not transfer well between skills. Contrastively, declarative knowledge is more likely to transfer, although there are also considerable differences between receptive and productive knowledge for L2 learners (e.g., measured by vocabulary size tests; see Webb, 2008). DeKeyser (2015) argued that the existence of the transfer between comprehension and production skills is limited and can be related to the fact that practice in either comprehension or production skills reinforces the declarative knowledge (e.g., vocabulary and grammar). There has been empirical evidence that input-based practice (e.g., reading newspapers and listening to podcasts) is more beneficial for developing L2 comprehension (listening and reading), with output-based practice (e.g., giving oral presentations and writing essays) being more beneficial for developing L2 production (speaking and writing) (see De Jong, 2005; DeKeyser & Sokalski, 2001; Rodgers, 2011). Since SA and AH learning contexts may differ in the extent to which they provide opportunities for input- and output-based practice, comprehension and production skills may benefit from studying abroad to different extents and therefore the evaluation of study abroad effects should take this distinction into account.

Moreover, language learning practice may also be modality-specific (auditory vs. visual modalities), such that learning in one modality may have limited transfer to that in another (see Robinson, 2003). For example, reading ability developed by processing visually-presented input (i.e., text) may not transfer well to listening ability that depends on processing of auditorily-presented input (i.e., speech), particularly for language learners coming from a completely different script (as is the case for Chinese learners of English; see Ma, Yu, & Zhang, 2018). Hence, an imbalance between reading and listening proficiency and development is common for foreign language learners in a formal education setting. As naturalistic immersion in SA learning contexts features large amounts of spoken interaction relative to formal instruction in AH learning contexts, it is conceivable that studying abroad might be more beneficial to the development of L2 listening and speaking than that of L2 reading and writing.

C. Type of measurement

Measures of language proficiency can be roughly categorized as general proficiency measures (e.g., oral proficiency and reading proficiency), declarative knowledge measures (e.g., vocabulary and grammar tests) and processing-related measures (e.g., oral fluency and spoken-language processing efficiency). According to Collentine (2004), compared to an AH context, an SA context is more effective in facilitating the development of fluency-related measures (i.e., narrative ability and semantic density) than that of discrete grammatical and lexical features in oral production. Our Chapter 3 showed that the SA learners improved more than their AH counterparts in speed of processing during listening comprehension, while the effect of learning context on auditory vocabulary acquisition could not be observed. These two studies seemed to suggest that the knowledge and processing aspects of language learning might not benefit from studying abroad to the same extent. This may be associated with the differences between formal education in AH contexts and naturalistic immersion in SA contexts. However, it is generally not yet clear whether and how the effects of studying abroad may differ between general proficiency, declarative knowledge, and processing-related measures.

D. Length of stay

Length of stay plays a critical role in SA research (DeKeyser, 2014). As language acquisition is a slow process and social integration for study-abroad learners also takes time, it is unclear how long an SA program should at least be for learners to achieve considerable language progress. Previous studies have shown mixed results on whether and to what extent short-term study-abroad programs, ranging from several weeks to a semester, were effective in facilitating language acquisition (e.g., Collentine, 2004; Llanes & Muñoz, 2009, 2013; Segalowitz & Freed, 2004; Serrano et al., 2011; Isabelli-Garcia, 2010). On the other hand, longer stays may not necessarily lead to greater language gains, with some studies providing positive evidence (e.g., Sasaki, 2011) and others negative evidence (e.g., Lara, Mora, & Pérez-Vidal, 2015). Thus, there seems to be no clear patterns as to whether and how length of stay may influence study abroad effects across studies. Note that previous meta-analyses (i.e., Yang, 2016; Varela, 2017; Xu, 2019) investigated the potential moderation effect of this variable, arriving at different conclusions. This will be elaborated upon in section 5.1.3.

E. Cultural distance

Cultural distance refers to differences between two cultures in terms of habits, values, and communication styles (Shenkar, 2012), which students studying abroad may experience. According to Maertz, Hassan, and Magnusson's (2009) cultural cognitive dissonance theory, the gaps between expatriates' native and host cultures create internal conflicts and discomfort for learners, which in turn compels them to learn and adjust. A new internal balance will be achieved when enough knowledge of the hosting culture (including language) has been acquired. Varela (2017) hypothesized a facilitative role of cultural distance in language learning but did not find empirical evidence for this in his meta-analysis. He argued that "the anxiety and frustration rooted in incongruent communication styles cancelled out the amplified opportunities that large cultural gaps might create". We are curious to see if this finding on cultural distance can be replicated in the present study, despite differences in the design of our meta-analysis.

5.1.3 Previous meta-analyses of SA effects

As summarized in Table 5.1, we have identified five previous meta-analyses on the effects of studying-abroad language learning (i.e., Tseng, Liu, Hsu, & Chu, 2021; Tullock & Ortega, 2017; Varela, 2017; Xu, 2019; Yang, 2016). The inclusion criteria of Tseng et al. (2021), Varela (2017) and Yang (2016) were comprehensive, including all measures that were indicative of L2 proficiency, while Tullock and Ortega (2017) and Xu (2019) focused on one aspect of L2 proficiency, i.e., oral fluency and complexity of language, respectively. Brief reviews of these previous meta-analyses are presented in order of publication date below. Note that effect sizes were evaluated based on Cohen's (1988) benchmarks of small (d = 0.2), medium (d = 0.5) and large (d =0.8).

Table 5.1: Summary of previous meta-analyses of SA effects on L2 learning.

Study	**Research focus**	**No. of included studies**	**Design of included studies**	**Pooled effect size and 95% CI**	**Significant moderators**
Yang (2016)	L2 proficiency	11	Between-group pretest-posttest design*	0.75 [0.56, 0.95]	Length of stay
Varela (2017)	L2 proficiency	33	Mixed	0.98 [0.78, 1.17]	Length of stay
Tullock & Ortega (2017)	L2 oral fluency	31	Within-group pretest-posttest design	Not applicable (effect sizes ranging from 0.5 to 1.2)	Not applicable
Xu (2019)	L2 complexity	28	Within-group pretest-posttest design	0.37 [0.24, 0.49]	Pre-departure proficiency level; type of coursework; language pledge**
Tseng et al. (2021)	L2 proficiency	42	Mixed (between-group)	0.87 [0.53,1.20]	Length of stay

Notes: **Although the primary studies of Yang's (2016) meta-analysis were of between-group pretest-posttest designs, the actual calculation of effect sizes was based on posttest performances only.* ***According to Xu (2019), a language pledge was a promise made by the student to speak only the target language while studying abroad.*

Yang (2016) was the first meta-analysis focusing on studies with a between-group pretest-posttest design. It quantitatively synthesized 11 studies comparing second language development in SA and AH learning contexts, which yielded a medium-to-large overall effect size favoring the SA learning context (d = 0.75). Yang reported that length of stay was a moderator of SA effects. However, somewhat counterintuitively, short-term SA programs (from 11 weeks to 13 weeks) were found to be more effective than long-term ones (from 16 weeks to 3.5 years) in facilitating L2 development. This meta-analysis was criticized by Xu (2019) for the inclusion of two particular studies with enormous effect sizes (i.e., d = 7.797 and 5.454), which might have skewed the results. Furthermore, it should be noted (as mentioned before) that Yang (2016) only extracted data from the posttest of individual studies for the calculation of effect sizes, without correcting for the potential between-group differences at pretest. This might have led to an undue inflation of effect sizes.

The meta-analysis by Varela (2017) included 33 individual studies, reporting a large effect size (d = 0.975) of studying abroad on second language acquisition. Note that this meta-analysis included studies with either within-group pretest-posttest designs or between-group posttest-only designs, which were not treated separately. Moreover, the author assigned a fixed value (i.e., 0.5) as pretest-posttest correlation coefficient whenever pretest-posttest correlations could not be retrieved, without conducting a sensitivity analysis over a range of possible values afterwards. Regarding moderator analyses, Varela found that length of stay moderated SA effects across studies, with longer stays being associated with larger effect sizes. Meanwhile, there were no significant moderation effects of cultural distance and type of immersion (i.e., homestay vs. dorm). Note that in Varela (2017) homestay was regarded as a more immersed context than dorm, as it provides more opportunities for socializing with locals.

Tullock and Ortega (2017) performed a meta-analysis of 31 studies that investigated the development of oral fluency in SA learners with a within-group pretest-posttest design. The authors chose not to aggregate effect sizes of individual studies into an overall effect size due to a lack of methodological consistency among included studies. They reported that individual effect sizes (indicated by Cohen's d) for speech rate ranged mostly from 0.5 to 1.2, and concluded conservatively that L2 learners' oral fluency probably improved after studying abroad. The authors warned that the magnitude of the fluency gains should be interpreted with caution because oral fluency was quantified differently across studies.

Xu (2019) meta-analyzed 28 studies that examined oral and written complexity in learners before and after an SA experience. Xu (2019) reported an overall small-to-medium effect (*d* = 0.37) of study abroad on the development of L2 complexity. Xu found that pre-departure proficiency level moderated the magnitude of the SA effect, with the intermediate pre-departure proficiency level associated with the largest SA effect sizes. Note, once again, that the individual studies included in Tullock and Ortega (2017) and Xu (2019) mostly adopted a within-group pretest-posttest design, such that there were no control groups to compare the development of the SA groups to.

Tseng et al. (2021) performed a multi-level meta-analysis of 42 primary studies adopting either a between-group pretest-posttest or posttest-only design. The results showed a medium-to-large effect (*g* = 0.87) favoring studying abroad over at home. Different from the univariate meta-analysis in other previous meta-analyses, this multi-level meta-analysis included all effect sizes for each measure used in the primary studies, allowing for a comprehensive investigation of the effects of moderator variables. The authors derived from their moderator analyses that learners of lower baseline proficiency, who also took both formal and content-based language courses and lived with host families during their stay abroad, tended to make more progress when studying abroad. Similar to Varela (2017), study-abroad learners' language gains were found to be positively associated with length of stay. Importantly, this meta-analysis by Tseng and colleagues did not provide details on the calculation of effect sizes, leaving it unclear whether or how different methodological designs (i.e., between-group pretest-posttest design and between-group posttest-only design) were taken into account when extracting effect sizes from primary studies.

These meta-analyses have provided estimations of weighted average SA effects across individual studies on oral fluency, language complexity, and/or general learning outcomes. They have also provided insights in factors that might moderate the magnitude of SA effects. At the same time, (the studies in) the meta-analyses were subject to one or more of the following limitations: lack of a comparison group, including a comparison group but not taking pretest between-group differences into account, mixing different methodological designs, or not including all measures of individual studies.

5.1.4 The current study

To further investigate the impact of studying abroad on second language acquisition, the present study quantitatively synthesized twenty studies, all having a between-group pretest-posttest design (i.e., comparing second language development in study-abroad and at-home learning contexts over a certain period of time). We fitted multilevel meta-analytical models to estimate the overall effect size of studying abroad and to explore possible moderating factors of studying-abroad effect.

The present meta-analysis focused solely on studies with a between-group pretest-posttest design for reasons of methodological rigor, as previous meta-analyses either lacked a comparison group or failed to control for between-group pre-departure differences. By adopting Morris' (2008) effect-size-calculation algorithms for pretest-posttest control designs, the present meta-analysis is able to compare between-group differences in terms of change from pre-test to post-test. Importantly, we also examined whether methodological design (i.e., between-group posttest-only design, within-group pretest-posttest design, between-group pretest-posttest design) modulated the size of SA effects through simulation analyses. Although our pool of included studies only represented between-group pretest-posttest design, we simulated the other two design categories by analyzing only partial data of our included studies (see section 5.2).

The multi-level meta-analysis we employed has certain advantages. As the included studies usually have multiple measures, we could include all measures which better represents the original studies than selecting only one measure per study. This approach also improves replicability of the meta-analyses, compared to univariate meta-analyses for which the selection criteria of a single measure per study may not be clear. Furthermore, inclusion of effect sizes for multiple (rather than single) measures per study allows for more fine-grained moderator analyses.

In sum, to synthesize existing SA research and to explore potential moderators of SA effects, we conducted a systematic review with a multi-level meta-analysis of studies comparing second language development in SA and AH learning contexts. Our research questions were as follows.

a. What is the overall effect of study-abroad learning contexts, in comparison to at-home learning contexts, on second language development between pretest and posttest?
b. Do study characteristics (i.e., methodological design, specific language skills in question, type of measurement, length of stay, cultural distance) influence the observed effect of studying abroad?

Note that the first research question will be address by a multi-level meta-analysis, and the second by analyses simulating different designs and by moderator analyses (see section 5.2.6 for details).

5.2 Methods

5.2.1 Literature search

We searched for empirical studies and review articles on study-abroad language learning in three electronic databases: Web of Science, Google Scholar, and Education Resources Information Center in September 2020. The search strategy was formulated as follows: (“stud* abroad” OR “immersion” OR “residence abroad” OR “exchange” OR “learning context” OR “learning environment”) AND “language” AND (“pre-post” OR “pretest” OR “posttest” OR “longitudinal”). In a next step, the included studies of previous meta-analyses were screened (i.e., Tseng et al., 2021; Tullock & Ortega, 2017; Varela, 2017; Xu, 2019; Yang, 2016), as well as the reference lists of these included studies.

5.2.2 Inclusion criteria

The included studies met the following criteria.

a. Apart from an SA target group, the studies had an AH comparison group receiving regular formal instruction in their home country.
b. The studies adopted a pre- and post-test design to demonstrate the development of second language proficiency over a period of time. Studies with multiple measurement times were also included, although only the first and the last times of measurement were taken into account.
c. The target measures of the studies were objective indicators of L2 proficiency. Studies with self-reported measures were therefore not included in this meta-analysis.
d. The participants of the studies were adult L2 learners.
e. All the studies were published, as "the inclusion of data from unpublished studies can itself introduce bias" (Higgins et al., 2019). This is because the unpublished studies that a meta-analyst is able to retrieve may not be representative of all the unpublished studies. Besides, unpublished studies have not undergone the methodological scrutiny of peer-review.
f. The studies were written in English.

The purpose of these criteria is to ensure the quality of the included studies and minimize confounds that might stem from differences in methodological design, measurement of proficiency, and participants' age. Note that studies that involved the same participants but employed different measures were all included, but they were labelled as one study in the analysis to avoid inflating the weight of individual studies.

5.2.3 Data coding

Five types of study characteristics were coded as potential moderators: length of stay, language dimension, language modality, type of measurement, and cultural distance. Length of stay was operationalized as a variable of two levels: "short-term stay" (shorter than or equal to 16 weeks) and "long-term stay" (longer than 16 weeks), following Xu (2019). Note that length of stay was not treated as a continuous variable because its distribution is not normal ($W = 0.80$, $p = < .001$). Language dimension was a variable of two levels: "comprehension" (referring to tasks tapping into listening and reading) and "production" (referring to tasks tapping into speaking and writing). Similarly, language modality had two levels: "auditory modality" (referring to tasks tapping into speaking and listening) and "visual modality" (referring to tasks tapping into reading and writing). Type of measurement assigned the measures of the included studies to one of the following three categories: 1) "general proficiency" (referring to holistic measures that assessed general proficiency), 2) "knowledge" (referring to discrete measures that assessed knowledge of specific aspects of language, e.g., grammar and vocabulary), and 3) “processing” (referring to processing-related measures, e.g., oral fluency, writing fluency, and spoken-language processing efficiency). Finally, following Varela’s (2017) practice, cultural distance was operationalized as a numeric variable based on Hofstede’s (2001) four cultural dimensions (i.e., power distance, individualism, masculinity, and uncertainty avoidance). It was a composite score, calculated by Kogut and Singh’s (1998) guidelines on Euclidian distance, indicating the differences between the native and the host country in the four cultural dimensions (see Varela, 2017 for a list of cultural distance scores, as also used in the present study).

Unlike univariate meta-analyses, multi-level meta-analysis allows for coding per measure (nested under study) instead of per study, resulting in a more comprehensive SA vs. AH group comparison and hence allowing more comprehensive moderator analyses. See Appendix Table G2 in the supplementary materials for our specific data coding.

5.2.4 Data extraction

For all included studies, the mean, standard deviation, and sample size of each group for each measure at both pretest and posttest, together with the pretest-posttest correlations for each group on all measures, are needed to calculate effect sizes of language proficiency gains achieved by each group over the respective study periods (see Morris, 2008; for the extracted data, see Appendix Table G3 in the supplementary materials). Firstly, we extracted the means, standard deviations, and sample sizes from the publications. In some cases, standard deviations were manually calculated as standard errors multiplied by the square root of samples sizes. When the primary studies reported general measures and fine-grained sub-category measures of the same data, we only extracted the statistics of the fine-grained measures to avoid confounded effect size estimates. The often unreported pretest-posttest correlations are critical to account for the dependency between pretest and posttest performance. The lack of pretest-posttest correlations is a common problem for meta-analysts dealing with repeated-measures designs, as it leads to imprecision in the confidence interval of the overall effect (Soveri, Antfolk, Karlsson, Salo, & Laine, 2017). A remedy to this imprecision problem, as suggested by Borenstein, Hedges, Higgins and Rothstein (2009), is to identify a plausible range for the correlation and then perform a sensitivity analysis. Therefore, when the correlations were not available, we set the pretest-posttest correlations at 0.5, which is the average of the four studies for which pretest-posttest correlations were made available (i.e., Jochum, 2014; Llanes & Muñoz, 2013; Serrano et al., 2011; Chapter 3 of the present thesis). Note that the correlation value of 0.5 was also used in Varela’s (2017) previous meta-analysis. Afterwards, a sensitivity analysis across the possible range of correlations (from 0 to 1) was conducted to determine the robustness of our meta-analysis results.

5.2.5 Calculation of effect sizes

The calculation of effect sizes can be divided into two steps. First, to calculate the effect size of pre-to-post changes for each participant group, we employed the *escalc* function in the Metafor package in R (Viechtbauer, 2010). Means, standard deviations, sample sizes, and pretest-posttest correlations were fitted into this function, which returned two types of values, namely the effect size and variance of the pre-to-post changes for each group (i.e. g_{within} and V_{within}). Second, after obtaining the within-group effect sizes and variances, we calculated the between-group differences (i.e., the differences between the SA and AH groups) in pre-to-post changes by subtracting the g_{within} of the AH group from that of the SA group or vice versa, depending on the directionality of the specific measures (i.e., for measures like accuracy or composition score, the larger values represents better performance; for error rate or reaction time, the smaller the better). The variance of the between-group differences was the sum of the variance for the SA group and that for the AH group (Viechtbauer, 2017). Additionally, we calculated the fail-safe N index (Rothstein, Sutton, & Borenstein, 2005) based on between-group effect sizes and variances, in order to check the likelihood of publication bias in our sample of studies. If the fail-safe number (i.e. N_{fs}) is greater than the safe number (i.e., $N_s = 5k + 10$, with k representing the number of effect sizes), the likelihood of publication bias would be minimal. Note that there are 105 effect sizes in the present study, so k equals 105 and N_s equals 535.

5.2.6 Multi-level meta-analytical modeling

The effect sizes and variances calculated above were fitted into a multi-level meta-analytical model to estimate the overall effect size and its confidence interval, through the *rma.mv* function in the Metafor package in R (Viechtbauer, 2010). This model took ‘study number’ and ‘measure number’ variables as random intercepts, estimating the variance of different studies as well as that of different measures within a study. Therefore, our multi-level random-effects model could account for the dependency of effect sizes between measures within studies. However, the dependency of effect sizes due to multiple between-group comparisons per measure within a study (i.e., the same SA group was compared to multiple AH groups or vice versa) could not be estimated adequately due to limited data. More specifically, only three out of the twenty included studies contained more than one SA vs. AH group comparisons (i.e., Sasaki, 2011; Serrano et al., 2011; Chapter 3 of the present thesis). In line with Soveri et al. (2017), we addressed this level of dependency beforehand by conducting internal fixed-effects meta-analyses on the effect sizes of the (dependent) multiple between-group comparisons per measure for the three aforementioned studies. This ensured that each study has only one SA vs. AH group comparison per measure, thus preventing dependencies between multiple between-group comparisons in our multilevel meta-analysis. Note that intensive domestic immersion groups were excluded from the comparisons (Segalowitz & Dewey, 2004). This is because intensive domestic immersion programs involve formal instruction given exclusively in the target language (albeit in the learners' home country) and numerous out-of-class learning activities with a communicative language focus (Freed et al., 2004). As such, such programs are considered an atypical AH learning context.

A sensitivity analysis was conducted to determine the robustness of our results across the possible range of pretest-posttest correlations. That is, we examined whether the overall effect size and its confidence interval would differ considerably across the possible range of pretest-posttest correlations between 0 and 1. If 0 was entered as a correlation, it meant that the pre- and post-test were treated as independent measurements, which in turn led to a possible inflation of the standard error of the pretest-posttest comparison and consequently reduced the weight in the meta-analysis. The opposite was true for entering 1 as a correlation, which strongly reduced the standard error of the pretest-posttest comparison and thus increased the weight in the meta-analysis.

Afterwards, simulation analyses were performed to determine whether different choices of methodological design would influence the magnitude of observed effect sizes. We simulated the types of designs we did not include in our meta-analysis, namely between-group posttest-only designs and within-group pretest-posttest designs, by analysing partial data of our included studies (of between-group pretest-posttest designs). More specifically, the between-group posttest-only designs were simulated by discarding the pretest data and retaining the posttest data, while the within-group pretest-posttest designs were simulated by excluding the AH groups and reserving the SA groups. Separate meta-analyses were then performed for these two types of designs.

Finally, we conducted a series of moderator analyses to investigate how study characteristics (i.e., language dimension, language modality, type of measurement, length of stay, cultural distance) might moderate the magnitude of SA effects on second language development. For categorical moderator variables, we examined whether the moderation effect reached significance and, if so, we investigated SA effects for each subcategory. Similarly, for the one numerical moderator variable (cultural distance), we first tested the significance of its moderation effect and then, if significant, examined the directionality of this possible moderation effect.

5.3 Results

5.3.1 Descriptive statistics

This study sample consisted of twenty studies, two of which shared the same participants. In total, 240 within-group effect sizes indicating the magnitude of language development over a certain period of time were derived for pretest and posttest performance of 44 groups (i.e., 22 SA and 22 AH groups). The number of within-group effect sizes was reduced to 210 after addressing the dependency of multiple group comparisons per measure within three studies. This yielded 105 between-group effect sizes (i.e., SA vs. AH group comparisons). These between-group effects sizes, as displayed in Figure 5.1, were entered the meta-analytical models to estimate the overall effect size and explore the possible moderation effects of several study characteristics. Appendix Table G2 presents more detailed information about within-group and between-group effect sizes of each individual study.

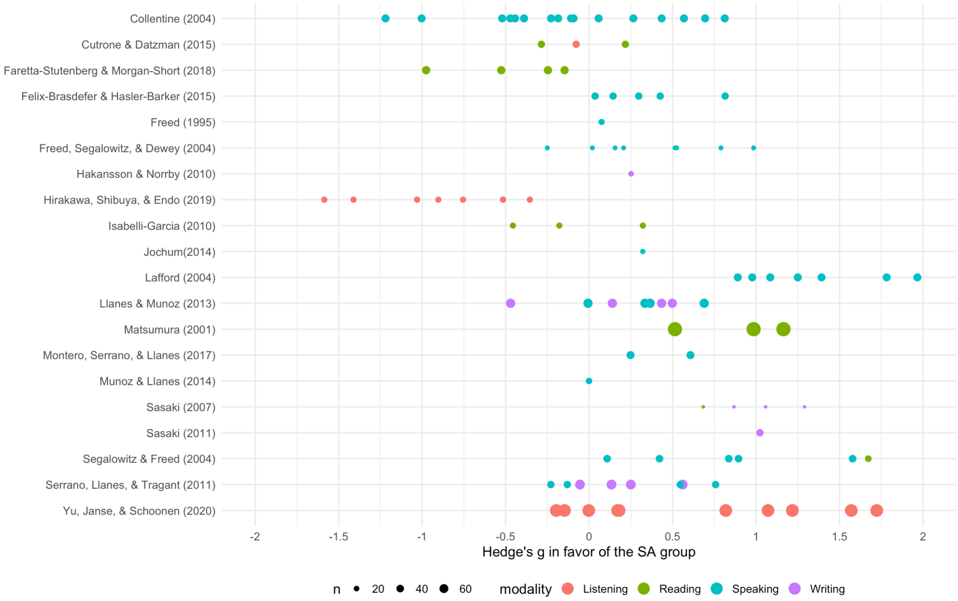

Figure 5.1: Effect sizes of the twenty included studies. A dot refers to a between-group effect size (i.e., SA vs. AH group) on a measure within a study. The size of a dot is proportional to the sample size of the concerned study. The color of a dot indicates which language skill (i.e., listening, reading, speaking and writing) the concerned measure tapped into. The horizontal axis is Hedge's g, with positive values favoring the SA groups and negative values favoring the AH groups. Five effect sizes are outside the plotted range, which can be retrieved in Appendix Table G2.

5.3.2 The overall effects of studying abroad

The multi-level analytical model estimated that the overall effect size of studying abroad on second language development is 0.31, with the 95% confidence interval ranging from 0.03 to 0.60. This is a small-to-medium effect, based on Cohen's (1988) benchmarks. The residual heterogeneity test was significant ($Q(104) = 593.77$, $p < .001$). In addition, the fail-safe N index analysis yielded a value of N_{fs} =4540, which far exceeds the critical value of N_s = 535 mentioned earlier. This fail-safe index thus demonstrates that the likelihood of potential publication bias influencing our meta-analytical results was minimal.

The sensitivity analysis across the possible range of pretest-posttest correlations showed that, when setting the unknown correlations to 0, the overall effect size is 0.32 [0.03, 0.61]. When setting the unknown correlations to 1, the overall effect size is 0.31 [0.02, 0.59]. Therefore, our results regarding the overall SA effects on second language development are robust, as the confidence intervals across these different correlations were similar and none of them included zero.

Simulation analyses showed that SA effects were the largest for between-group posttest-only designs (*g* = 0.62, [0.11, 1.13], *p* = .017), followed by within-group pretest-posttest designs (*g* = 0.40, [0.25, 0.54], *p* = < .001), and were the smallest for the between-group pretest-posttest designs (*g* = 0.31, [0.03, 0.60], *p* = .032) (see Table 5.2).

Table 5.2: Estimates of overall study-abroad effects for different methodological designs.

	β	*SE*	*z*	*p*	*95% CI*
Between-group pretest-posttest design	0.31	0.15	2.14	.032	[0.03, 0.60]
Within-group pretest-posttest design	0.40	0.07	5.30	< .001	[0.25, 0.54]
Between-group posttest-only design	0.62	0.26	2.38	.017	[0.11, 1.13]

5.3.3 Moderator analyses

Firstly, we tested whether the observed SA effects were moderated by the variable language dimension (i.e., comprehension and production). The moderation effect of language dimension on SA effects was found to be only marginally significant ($Q_M(2)$ = 5.33, *p* = .070). As shown in Figure 5.2, the synthesized SA effect for comprehension failed to reach significance (*g* = 0.19, [-0.23, 0.61], *p* = .376), but that for production was significant (*g* = 0.39, [0.05, 0.73], *p* = .026).

We then investigated SA effects by language modalities (auditory vs. visual), but did not observe a significant moderation effect of language modality ($Q_M(2)$ = 4.57, *p* = .102). Hence language modality did not moderate the observed SA effects in this meta-analysis.

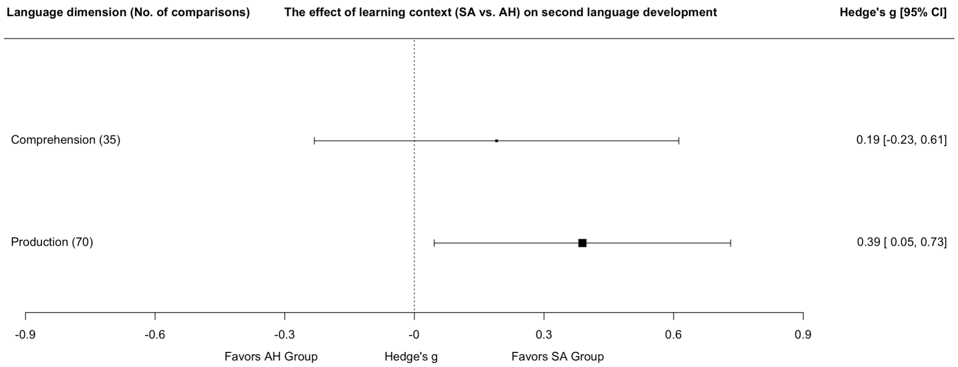

Figure 5.2: Synthesized SA effects by language dimensions (comprehension and production).

Length of stay moderated the magnitude of SA effects ($Q_M(2) = 8.11$, $p = .017$). As shown in Figure 5.3, the synthesized SA effect was significant for long-term stay (g = 0.76, [0.19, 1.33], $p = .009$), but not significant for short-term stay (g = 0.18, [-0.13, 0.49], $p = .261$). Thus, the SA effect for long-term stay abroad was significantly larger than that for short-term stay abroad.

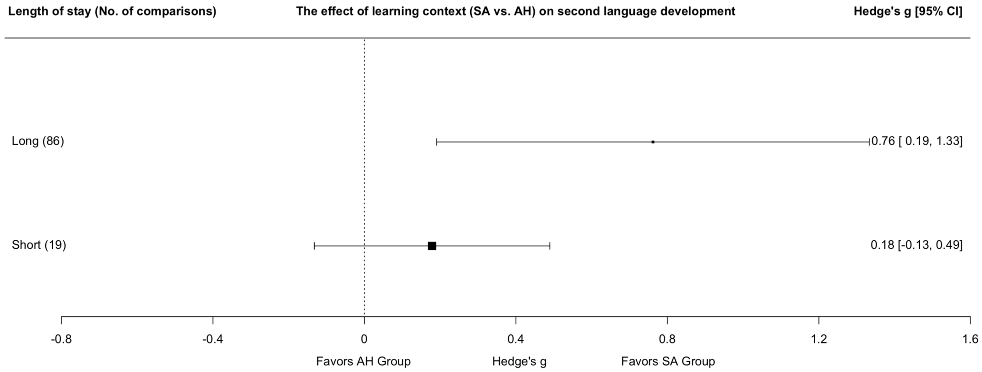

Figure 5.3: Synthesized SA effects by length of stay (long- and short-term).

We also found that type of measurement was a significant moderator of SA effects ($Q_M(3) = 45.71$, $p < .0001$). As shown in Figure 5.4, SA effects were the largest for general proficiency measures (g = 0.66, [0.33, 1.00], $p < .001$), followed by processing-related measures (g = 0.39, [0.07, 0.72], $p = .017$), while SA effects for knowledge measures were in the opposite direction (g = -0.53, [-0.91, -0.16], $p = .005$).

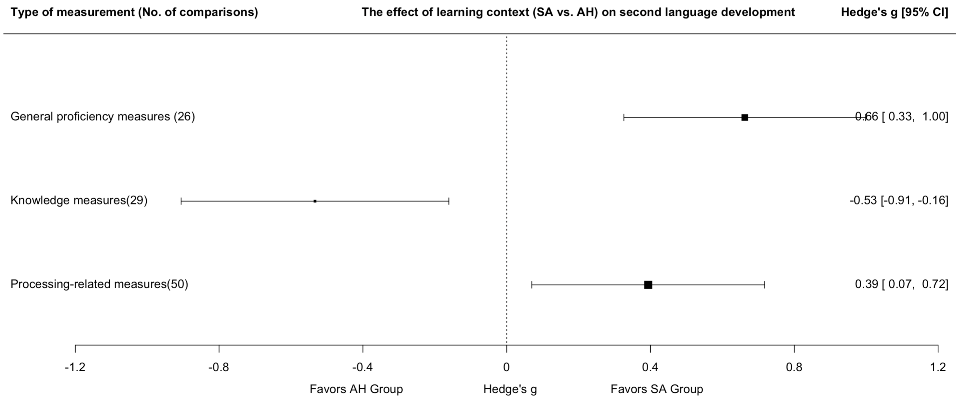

Figure 5.4: Synthesized SA effects by measurement type (general proficiency, knowledge, and processing).

Finally, we tested the moderation effect of cultural distance (a numerical variable). The results showed that cultural distance did not moderate SA effects($Q_M(1) = 0.001$, *p = .979*).

5.4 Discussion

Our first research question concerned the overall effect of study-abroad learning context on second language development in comparison to at-home learning context. The meta-analysis revealed a small-to-medium overall effect of learning context (i.e., SA vs. AH) on facilitating second language acquisition ($g = 0.31$). This result suggests that study-abroad learners indeed benefited from an immersive learning environment characterized by higher amount and quality of target language exposure, which agrees with the common belief about the benefits of studying abroad for language learning. There are at least two possible explanations for the modesty of the overall effect size. Firstly, the degree of immersion, which depends heavily on social integration and acculturation, may not be as high as expected for newly-arrived study-abroad learners (see Coleman, 2013). Hence study-abroad learners may not necessarily enjoy the supposedly high amount and quality of target language exposure associated with full immersion.-Secondly, the study-abroad programs in question were usually shorter than a year. Programs of longer durations might induce larger SA effects. On a related note, the fact that the overall SA effect was rather small may, to some extent, account for some researchers' skepticism about the linguistic benefits of SA (e.g., DeKeyser, 2010; Rees & Klapper, 2008). Compared with Tseng et al.'s (2021), Yang's (2016) and Varela's (2017) meta-analyses, all of which reported a large or medium-to-large overall effect of studying abroad, the present meta-analysis offered a relatively conservative estimation of the overall SA effect. The differences between the presently reported size of the SA effect and those reported in these three previous meta-analyses are likely to stem mainly from differences in methodological design of included studies (inclusion of comparison groups or correction for pre-departure group differences).

Our second research question addressed the moderating factors of study-abroad effects reported by the included studies. It was found through simulation that methodological design modulated the magnitude of observed SA effects. Moreover, two moderators of SA effects were found significant in moderator analyses: length of stay and type of measurement. The moderation effect of language dimension was marginally significant. No moderation effects of language modality and cultural distance were observed.

Firstly, through simulating other methodological designs, we found that between-group posttest-only designs manifested the largest SA effects, followed by within-group pretest-posttest designs and between-group pretest-posttest designs, in that order. The reason why between-group posttest-only designs yielded the largest study-abroad effects is probably because between-group differences before studying abroad were not accounted for in this design. Since SA learners may have already started spending more time and effort learning the target language, compared to their at-home counterparts, to prepare themselves for studying abroad, it is likely that SA learners were more proficient in the target language than their at-home counterparts before going abroad (see section 5.1 for consequences of non-randomized group assignment). That is, this design may unduly inflate SA effect sizes. Therefore, the present study provides empirical evidence for Tseng et al.'s (2021) and Tullock and Ortega's (2017) argument that between-group posttest-only designs are not desirable for SA research. Within-group pretest-posttest designs yielded larger SA effects than between-group pretest-posttest designs because the former measured absolute gains of SA learners while the latter measured comparative gains (i.e., SA vs. AH). The choice between these two designs should obviously depend on the research questions of a specific study. When the research focuses on the linguistic benefits of studying abroad, between-group pretest-posttest designs are more appropriate as they can rule out the confound of general improvement irrespective of learning context (see Cook & Campbell, 1979).

Secondly, the moderator analysis on length of stay showed that studies with long-term SA experiences (i.e., longer than one semester) demonstrated larger effect sizes than those studies with short-term SA experiences (i.e., shorter than or equal to one semester). This means that long-term SA experiences were more effective in facilitating second language development than short-term SA experiences. This may be related to the slow pace of the socialization process for international students during their studying abroad (Coleman, 2013), which conditions the number of opportunities for oral interactions in the target language. The greater benefits of longer-term stays may also be related to the accumulative nature of language acquisition. This finding for length of stay aligns with Tseng et al.'s (2021) and Varela's (2017) meta-analyses, but not with Yang's (2016) or Xu's (2019) meta-analyses. Yang (2016) reported the opposite result that short-term SA was actually more effective than long-term SA, but warned to interpret this finding with caution. This unexpected finding of Yang (2016) on length of stay can be accounted for in two ways. Firstly, this finding may, to some extent, have been driven by the inclusion of two studies with enormous effect sizes (as mentioned before in section 5.1), which were both classified as short-term studies. As already mentioned in the *Introduction*, inclusion of these two studies may have inflated the effect size of this category. Secondly, as pointed out by Yang (2016), there might be a potential confounding factor (i.e., the multidimensional aspects of language testing) as there was a wide range of variation in effect sizes and the magnitude of effect sizes corresponded to how specific the assessed skills were. Hence, Yang suggested that the reason why short-term studies were found more effective than long-term studies may be associated with that the former assessed more specific skills than the latter. Xu (2019), on the other hand, showed short-term and long-term SA programs did not differ in facilitating gains on L2 complexity. As measures of language complexity, especially syntactical complexity, have often been shown to be less sensitive to the effect of studying abroad than other measures (see e.g., Collentine, 2004; Llanes & Muñoz, 2013; Sasaki, 2007), the difference between the present meta-analysis and Xu's in the moderation effect of length of stay might be ascribed to the difference between the two in dependent variables concerned (i.e., L2 complexity vs. overall language proficiency).

Thirdly, the moderator analysis on type of measurement indicated that SA learners made more progress than AH learners on general proficiency and processing-related measures, but less progress on knowledge measures. This suggests that the advantage of an SA learning context in comparison to an AH learning context lies more in facilitating the acquisition of general proficiency and/or language processing than in facilitating acquisition of declarative language knowledge. This may be associated with the nature and purpose of language learning in these two learning contexts, with the SA context being more usage-oriented and the AH context being more exam-oriented. For SA learners, the target language functions as a carrier of information that has to be processed in real time for communication purposes, while for AH learners, it is often regarded as a complex system of knowledge that has to be acquired in order to prove L2 proficiency at exams. Therefore, the acquisition of knowledge and processing efficiency might be prioritized differently, depending on specific learning contexts.

This finding builds on and extends the results of previous studies that investigated the acquisition of knowledge and processing-related measures in the two learning contexts. More specifically, SA learners were usually found to show greater oral fluency than AH learners (Segalowitz & Freed, 2004; Freed, Segalowitz, & Dewey, 2004), while AH learners may show better development in the discrete lexical and grammatical features of their oral production (Collentine, 2004). As for listening, studying abroad may accelerate the acquisition of spoken-language processing, but not necessarily that of auditory vocabulary knowledge (see Chapter 3). As for writing, benefits of studying abroad were often reported for writing fluency, but not always for complexity measures of written production (Knoch, Rouhshda, Oon, & Storch, 2015; Sasaki, 2007). Therefore, it seems that the acquisition of knowledge and processing may show different developmental trajectories, and that they may not be equally sensitive to the effect of studying abroad.

Fourthly, the moderation effect of language dimension just missed significance (*p* = .070). Having a closer look at the different language dimensions, we found that SA effects were significant for L2 production but not for comprehension. It seems to suggest that compared to an AH learning context, the SA learning context tends to be more beneficial for the development of production skills (i.e., speaking and writing) than for that of comprehension skills (i.e., listening and reading). This may be because an SA learning context may entail more output-based learning practice than an AH learning context that is primarily limited to input-based learning practice. This finding thus seems to point in the direction of skill specificity, which claims that input-based practice facilitates comprehension skills and output-based practice facilitates production skills (DeKeyser, 2015), but we have to bear in mind that this moderation effect was only marginally significant. Alternatively, any potential difference between comprehension and production regarding the magnitude of the SA effect might also be related to measurement sensitivity. Measures for language production may be more sensitive than those for language comprehension as the former are usually open-ended and the latter (mostly in the form of multiple choice, matching, and cloze tasks) may suffer from ceiling effects. Future studies are clearly needed to further investigate the possibility of skill-specific effects of study-abroad language learning.

Fifthly, no moderation effect of modality was observed in this meta-analysis. Hence, we did not provide evidence for the belief that study abroad is more beneficial for skills pertaining to the auditory modality (i.e., speaking and listening) than for skills pertaining to the visual modality (i.e., reading and writing) (see Cubillos & Ilvento, 2013). This is may be because the study-abroad programs in question were likely to capitalize not only on listening and speaking but also on reading and writing. As formal education entails considerable amount of reading and writing activities (e.g., reading assignments, essay writing, and researching), it may not be justified to assume that SA provides more auditory input than visual input.

Finally, we did not observe a moderation effect of cultural distance, in line with Varela's (2017) meta-analysis. Hence we did not find evidence that cultural distance played a role for language learning abroad. This may be because cultural distance itself may have advantages and disadvantages for language learning abroad. On the one hand, cultural distance creates internal conflicts and psychological discomfort in international sojourners, which propels them to learn and adjust (Maertz, Hassan, & Magnusson, 2009; Cable, Gino, & Staats, 2013). The more distant between the original culture and the target culture, the stronger desire for adjustment participants might tend to experience. On the other hand, culture distance also creates barriers in communication, triggering anxiety and possibly even social withdrawal. As such, the facilitative and disruptive effects of cultural distance on language learning abroad might have indeed canceled each other out, as already argued by Varela (2017). Importantly, however, there was little diversity in the included studies in cultural distance, with over half of the studies investigating either English learners of Spanish or Spanish learners of English. Future studies might want to further investigate the role of cultural distance in a broader context.

One limitation of this study is that we did not study predeparture proficiency level as a potential moderator of SA effects, since there seemed to be evidence that predeparture proficiency level moderated SA effects in previous research (see section 5.1 for the summary of Xu's (2019) findings). However, as participants' predeparture proficiency in our included studies were mostly at an intermediate-level with little variance, the present study cannot sensibly investigate the possible moderation effect of this factor. However, the same holds for any meta-analysis of SA studies that mostly involve highly-educated university students with minimally intermediate levels of the target language. Another limitation of this study is the missing information about pretest-posttest correlations for the majority of the included studies, which affected the precision of our estimated SA effect. However, we have conducted a sensitivity analysis across the entire range of possible pretest-posttest correlations to ensure the reliability of our results. Nevertheless, we recommend that future study abroad studies report actual pretest-posttest correlations, which will improve the precision of future meta-analyses.

5.5 Conclusions

This study systematically reviewed the existing comparative studies of second language development in SA and AH learning contexts, in order to gain a comprehensive understanding of the impact of studying abroad on language learning. A multi-level meta-analysis revealed a small-to-medium overall effect of studying abroad on second language development (g = 0.31), which is smaller than in previous meta-analyses. This smaller effect may be primarily due to differences in methodological design of included studies. Follow-up analyses on the size of SA effects showed the following results. Firstly, between-group pretest-posttest designs provided the most conservative estimation of study-abroad effects, followed by within-group pretest-posttest designs, while between-group posttest-only designs considerably inflated study-abroad effects. Secondly, long-term study-abroad programs were more effective in facilitating L2 development than short-term ones. Thirdly, compared to their at-home counterparts, study-abroad learners showed more progress on general proficiency or processing-related measures than on declarative knowledge measures. Furthermore, although we found that SA effects were significant for L2 production but not for comprehension, the moderation effect of language dimension just missed significance (p = .070). Moreover, no significant moderation effects of language modality and cultural distance were observed. Altogether, this study has provided collective evidence on the linguistic benefits of studying abroad, as well as on several moderating factors of studying abroad effects, by quantitively synthesizing previous SA research.

Chapter 6: General discussion and conclusions

The aim of the present thesis was to study second language development and its individual variability in various learning contexts, with a particular emphasis on study-abroad language learning. Studying abroad has often been considered as the best second-language learning context, as learners have to inhibit their first language and immerse themselves in the second language (Freed, 1995; Jacobs, Fricke, & Kroll, 2016; Linck, Kroll, & Sunderman, 2009). At-home learning contexts, also referred to as foreign-language contexts, may be criticized for relatively inadequate exposure to the target language (henceforth: L2), over-reliance on rote-learning, and limited opportunity for interactive communication. Differences in learning contexts may result in different characteristics and trajectories of second language development (Kroll, Dussias, & Bajo, 2018), and may even affect how individual-difference factors (e.g., working memory) are associated with L2 learning processes and outcomes (Faretta-Stutenberg & Morgan-Short, 2018; Sunderman & Kroll, 2009). To investigate these claims about learning contexts, individual-difference factors, and L2 learning, I addressed the following research questions in this thesis.

1) Do English-as-a-foreign-language (EFL) learners from three different learning contexts (i.e., AH-regular, SA-onset, AH-intensive) differ in their L2 listening proficiency? How do they compare to native listeners? (Chapter 2)

2) Do study-abroad learners show more improvement in L2 listening proficiency than at-home learners over the course of an academic year? (Chapter 3)

3) What individual-difference factors are associated with listening proficiency and proficiency development across different learning contexts? (Chapter 4)

4) What is the overall effect of studying-abroad learning contexts, in comparison to at-home learning contexts, on second language development, based on results reported in existing study-abroad research? (Chapter 5)

Chapters 2 through 4 presented empirical investigations that zoomed in on listening comprehension, the least researched area in the field of second language acquisition research. The few listening studies that exist usually measured L2 listening proficiency in a holistic manner (e.g., with spoken passage comprehension tasks). The current study decomposed listening proficiency in light of skill-based theories (Anderson, 1983; DeKeyser, 2015) and theories of listening comprehension (Anderson, 2015; Goss, 1982; Cutler & Clifton, 1999). More specifically, L2 listening proficiency was operationalized in this thesis as listening vocabulary knowledge and spoken-language processing efficiency. In Chapter 5, we zoomed out to investigate the effect of study abroad on second-language development in general, by conducting a systematic review of existing study-abroad research.

The remainder of the current chapter starts with a summary of findings reported in Chapters 2 to 5 (section 6.1), and then relates these findings to current theories of second language acquisition (section 6.2). Methodological concerns, limitations, and future directions will be discussed next (section 6.3). Finally, I will bring together all the insights concerning the effect of study abroad on second language development from the separate chapters (section 6.4), in order to provide readers with a clear picture of what can be reasonably expected from short-term study-abroad programs.

6.1 Summary of findings

6.1.1 The effects of learning context on L2 listening development

Chapter 2 investigated whether learners from three EFL learning contexts differ in knowledge and processing aspects of L2 listening proficiency. Three groups of Chinese postgraduates, together with a control group of native speakers of English, took a battery of listening tests. The Chinese groups included non-English majors studying in China (AH-regular group), non-English majors newly arrived in the UK (SA-onset group), and English majors studying in China (AH-intensive group). Based on hours of instruction and self-directed learning activities, we assumed that the three EFL learning contexts differed with regard to amount of L2 exposure: the AH-intensive group having the highest exposure, the AH-regular group having the lowest exposure, and the SA-onset group being somewhere in between these two domestic groups (due to their preparation for studying abroad). The listening tests measured auditory vocabulary knowledge and listening processing efficiency (i.e., accuracy, speed and stability of L2 speech processing) in lexical access, grammatical processing, and semantic processing.

There were three major findings. Firstly, none of the three nonnative groups reached the level of the native group: neither in terms of vocabulary size, nor in terms of accuracy, speed, and stability of processing. The nonnatives and the natives were not close in their average performance on any of these listening tests, suggesting that these nonnative groups were far from ceiling performance. Secondly, the AH-intensive group was found to have a larger vocabulary size than the SA-onset group, who in turn outperformed the AH-regular group. The pattern of decreasing vocabulary sizes among these groups was likely due to their decreasing exposure levels conditioned by learning contexts. Thirdly, the AH-intensive and SA-onset groups did not differ in any of the processing measures, but they both outperformed the AH-regular group in accuracy and speed of processing across the three processing tasks. Hence these results seem to suggest that more EFL exposure is associated with larger vocabulary size but not necessarily with higher spoken-language processing efficiency. Note that the three nonnative groups did not differ in processing stability (measured by coefficient of variation). This suggests that these between-group differences mainly reflect quantitative differences in L2 cognitive processes but not qualitive differences (i.e., restructuring of cognitive processes) (see Segalowitz, 2010). Taken together, knowledge and processing aspects of L2 listening proficiency differed between groups and may not be equally affected by EFL exposure, which seems to have limited effect on facilitating language processing skills for intermediate-to-advanced learners. Note that these results also confirmed our speculation that the SA and AH-regular groups were not equal in L2 performance, although both were non-English majors and have been conventionally treated as comparable or proficiency-matched in study-abroad studies.

Chapter 3 investigated the effect of studying abroad on L2 listening development. This chapter built on the previous chapter in that the participants of Chapter 2 were invited to take the same tests again after roughly one academic year. This enabled us to compare L2 listening development in terms of language knowledge and processing efficiency in one study-abroad (SA) context and two at-home (AH) contexts over the course of one academic year. Note that the SA-onset group in Chapter 2 was relabeled as the SA group in Chapter 3.

Regarding vocabulary development, we found that the SA group made more progress than the AH-intensive group, but did not differ significantly from the AH-regular group. Firstly, as the AH-intensive group had the largest vocabulary size at pretest and did not show any improvement at posttest, this group might have reached a plateau. Hence, the difference between the SA and the AH-intensive cannot confirm our expectation of the facilitative role of an SA context in vocabulary acquisition (relative to an AH context). Secondly, the expected differences in vocabulary acquisition between the SA and the AH-regular group were not observed. This null effect has to be interpreted with caution, as it is unclear to what extent differences in baseline vocabulary size may have affected this between-group comparison of vocabulary growth. In other words, vocabulary growth may interact with pre-test scores, which in our data is hard to disentangle from the effect of learning context. However, the results seem to suggest that the magnitude of study-abroad benefits on vocabulary improvement may be small or non-existent. This study added to a series of previous studies that presented a mixed picture of SA effects on vocabulary development (e.g., Briggs, 2015; DeKeyser, 1991; Dewey, 2004, 2008; Ife, Vives Boix, & Meara, 2000; Llanes & Muñoz, 2009; Milton & Meara, 1995).

Regarding the acquisition of L2 processing efficiency, we found that the SA group made more progress than the AH-intensive group, but less progress than the AH-regular group, in speed of processing across tasks. Firstly, since the SA and AH-intensive groups started off with equal processing efficiency levels at pre-test, we argue that their different developmental patterns can be attributed to the effect of learning context. Secondly, as steep learning curves for low-proficiency participants in reaction time have been commonly observed in experimental settings, we argued that the AH-regular group made more improvement than the SA group due to its slower processing speed at baseline. Note that the three groups made equal progress in accuracy and stability of processing. These results suggest that, provided equal starting levels, studying abroad is more beneficial for enhancing L2 processing speed over remaining at home. This agreed with and extended existing literature concerning study-abroad effects on language processing, which focused on oral and written fluency (e.g., Freed, Segalowitz & Dewey, 2004; Sasaki, 2007; Segalowitz & Freed, 2004). Taken together, Chapter 3 demonstrated that study abroad may be an effective intervention for the acquisition of L2 processing efficiency (speed in particular), but not necessarily for vocabulary acquisition.

To summarize, Chapter 2 compared different EFL learning contexts (also referred to as at-home learning contexts), while Chapter 3 compared a study-abroad learning context against two at-home learning contexts. The results of these two empirical chapters suggest that the development of knowledge and processing aspects of L2 listening proficiency is affected by their specific learning context.

6.1.2 The effects of individual-difference variables on L2 listening development

Since learning context and individual-difference variables both affect the various aspects of L2 learning, understanding variability in L2 development requires joint investigation of these factors with careful consideration of their potential interplay (see Dörnyei, 2009; DeKeyser, 2012; Faretta-Stutenberg & Morgan-Short, 2018; Sanz, 2005). Previous studies on the interaction between individual-difference variables and learning context are scarce and have reported mixed results (e.g., Sunderman & Kroll, 2009; Tokowicz et al., 2004; Faretta-Stutenberg & Morgan-Short, 2018).

Chapter 4 aimed to examine the effects of several individual-difference factors on L2 listening comprehension in study-abroad (SA) and at-home (AH) learning contexts, probing into the possible interplay between individual-difference factors and learning context. This chapter built on the previous chapter in that the same participants were required to take individual-difference tests and questionnaires at the time of the posttest. As L2 learners may acquire language at different paces due to individual differences in language aptitude, working memory, mental well-being, language exposure, and social interaction, we tested how well these individual-difference factors predicted participants' pre- and post-test performance and their improvement across learning contexts.

Firstly, aptitude was found to positively correlate with participants' listening vocabulary size and their processing efficiency, more specifically, processing accuracy in all three processing tasks and processing speed in two out of the three tasks. This suggests that aptitude does not only relate to measures of crystallized language proficiency (including vocabulary size), but also to the efficiency with which the L2 is processed. Secondly, working memory did not relate to any of the listening measures. This is likely due to aptitude and working memory being both included in the statistical models, with the two having a considerable amount of shared predicted variance (see Mackey, Adams, Stafford, & Winke, 2010). Thirdly, language exposure (quantified as frequency of conducting language learning activities) was associated with processing efficiency. This association can be explained by the central concept of the skill acquisition theory of second language acquisition, stating that reaction time and error rate decrease with practice (DeKeyser, 2015). Note that we consider "language exposure", which is a term commonly used in usage-based theories, and "practice", which is a term commonly used in skill acquisition theories, to have a similar effect on second language learning in the present context. Fourthly, the amount of social interaction was not significantly related to any of the listening measures, suggesting that L2 learners with higher amounts of social interaction do not necessarily have higher listening proficiency. The null effect of social interaction has to be interpreted with caution, for this variable, operationalized as participants' estimated speaking time with others, may be difficult for participants to self-report and hence may not be valid. Fifthly, mental well-being did not relate to any of the listening measures, either. This suggests that mental well-being was either not relevant for adult L2 learning or participants might have not revealed their personal feelings in the questionnaires. Finally, unexpectedly, we found no interaction between the effects of individual-difference factors and learning context, meaning that learning context did not mediate the way individual-difference factors related to listening proficiency or its progress.

To summarize, Chapter 4 showed that, although none of the individual-difference factors predicted development over the academic year, some (i.e., aptitude and exposure) were stable predictors of listening proficiency at both pre- and post-test across learning contexts. Second language learning thus seems to be similarly related to individuals' capacities across different learning contexts.

6.1.3 Study-abroad effects and its moderators in previous research

Chapter 5 aimed to synthesize existing study-abroad research in order to estimate an overall study-abroad effect on L2 development (as compared to remaining in the home country) and identify factors that influenced the magnitude of observed study-abroad effects in individual studies. Twenty studies evaluating the efficacy of study-abroad programs (usually shorter than one year) in facilitating L2 development for university-level studies were included in our systematic review. These individual studies often revealed mixed and sometimes contradictory results, but a multilevel meta-analysis revealed a small-to-medium overall effect of studying abroad on L2 development in comparison to staying in the home country ($g = .31$). Compared to previous meta-analyses (i.e., Tseng et al, 2021; Yang, 2016; Varela, 2017), our meta-analysis yielded a conservative estimation of the overall SA effect size. The discrepancies are likely to stem mainly from differences in methodological design of included studies (inclusion of comparison groups or correction for pre-departure between-group differences). Nevertheless, this result suggests that study-abroad learners indeed benefited from an immersive learning environment characterized by higher amount and quality of target language exposure, which agrees with the common belief about the benefits of studying abroad for language learning. Furthermore, simulation analysis revealed that the between-group pretest-posttest design provided the most conservative estimation of study-abroad effects, followed by the within-group pretest-posttest design, followed, in turn, by the between-group posttest-only design. The between-group pretest-posttest design seems the most appropriate research design as it can account for general improvement irrespective of learning context as well as pre-departure between-group differences. Nevertheless, it is usually not feasible for researchers to randomly assign participants to groups in these kinds of field settings, so we still have to count with differences between internationally oriented students and students who prefer to stay at home (see Cook & Campbell, 1979).

Moderator analyses showed that studies with long-term SA experiences (i.e., longer than one semester) demonstrated larger effect sizes than studies with short-term SA experiences (i.e., shorter than or equal to one semester). This suggests that long-term study-abroad programs are more effective in facilitating L2 development than short-term ones. Moreover, compared to their at-home counterparts, study-abroad learners showed more progress on general proficiency and processing-related measures than on knowledge measures (mostly during the course of one semester or one year abroad). This finding agrees with Chapter 3, as well as other previous studies that investigated the acquisition of knowledge and processing-related measures in the two learning contexts (e.g., Segalowitz & Freed, 2004; Freed, Segalowitz, & Dewey, 2004; Collentine, 2004; Knoch, Rouhshda, Oon, & Storch, 2015; Sasaki, 2007).

To summarize, Chapter 5 reported a small-to-medium overall effect of studying abroad on second language development, and identified several factors (i.e., methodological design, length of stay, and type of measurement) that modulated the magnitude of observed SA effects.

6.2 Theoretical implications

The empirical studies in the current thesis investigated knowledge and processing aspects of listening proficiency, based on the skill acquisition theory of second language acquisition (DeKeyser, 2015). This theoretical framework distinguishes declarative knowledge and procedural knowledge. Note that processing efficiency, a key concept of this thesis, is one aspect and indicator of procedural knowledge. According to this skill acquisition theory of SLA, second language learning starts by explicit learning of declarative L2 knowledge. This knowledge then gradually becomes proceduralized and eventually may get automatized through practice (DeKeyser, 2015). DeKeyser (2015) stressed that this does not mean that declarative knowledge is to be transformed into procedural knowledge or that the more procedural knowledge, the less declarative knowledge. Literature on first language acquisition has also shown that infants with faster language processing tend to learn vocabulary more rapidly (see e.g., Peter, Durrant, Jessop, Bidgood, Pine, & Rowland, 2019). However, the developmental trajectories of knowledge and processing efficiency and their relationship for adult second language learners are unclear and understudied.

With respect to second language learning in an EFL learning context, we suggested that knowledge and processing aspects of listening proficiency are not equally sensitive to EFL exposure. EFL learning contexts seemed to be limited in the extent to which they facilitate processing efficiency (see Chapter 2). The empirical and meta-analysis results in this thesis have also shown that knowledge and processing aspects of L2 proficiency may benefit from studying abroad to different extents (see Chapters 3 and 5). The theoretical implication of these results is that the developmental trajectories for knowledge accumulation and processing automatization are likely to differ and they may be differentially affected by different learning contexts. Future studies are encouraged to further investigate the relationship between knowledge accumulation and processing automatization in the various learning contexts of adult second language acquisition and the factors that mediate the development of these two acquisition processes.

Regarding the proceduralization and automatization stages of second language acquisition, DeKeyser (2015) argued that the proceduralization of knowledge is not particularly time-consuming and may be complete after just a few trials/instances. Automatization of knowledge, however, may take much longer as learners may need to go through a large amount of practice to decrease the reaction time, error rates, and the amount of attention required when performing a linguistic task. The fact that automatization is a slow process has its repercussion in our studies. It is usually difficult to observe considerable progress in processing efficiency in foreign language classroom settings, but we set out to test whether an academic year abroad would be enough to facilitate observable improvement in processing efficiency relative to staying in the home country. We did observe a study-abroad effect on processing speed, but not on processing stability. One possible explanation for the lack of between-group differences in processing stability in our study is that the coefficient of variation (CV) may not be a sensitive measure to differentiate the nonnative groups in this study (see section 6.3). Alternatively, the results may be taken to indicate that a year abroad may not facilitate qualitative changes in L2 processing mechanisms. According to Segalowitz (2010), processing speed getting faster (i.e., across-the-board speeding up) indicates quantitative changes in language processing mechanisms, while greater processing stability indicates qualitative changes (i.e., a restructuring of processes). Therefore, our finding seems to suggest that a year abroad may facilitate quantitative changes but not qualitative changes in L2 processing mechanisms. Furthermore, long-term study-abroad programs were found to be more effective in facilitating second language development than short-term ones (Chapter 5). As the interval between pre- and post-test in our study (Chapter 3 and 4) is only one academic year, future studies are encouraged to investigate if longer study-abroad experiences could trigger qualitative changes in the L2 processing mechanisms of adult learners.

Another interesting but less researched aspect of DeKeyser's (2015) skill-based accounts of second language acquisition is skill specificity. Practice in one skill (e.g., listening) contributes to the improvement of this specific skill but not necessarily to the improvement of other seemingly related skills (e.g., speaking) (DeKeyser, 2007). Although declarative knowledge can be shared across the different language subskills, procedural knowledge is highly specific and may not be shared easily (DeKeyser, 2007). Hence, procedural knowledge for each skill does not transfer well between the language subskills. This thesis tapped into skill specificity in three ways, as elaborated below.

Firstly, we found that the social interaction variable (i.e., time of oral interaction) did not correlate with any of the listening measures (Chapter 4). This meant that L2 learners with higher amounts of social interaction did not necessarily have higher listening proficiency. The lack of association between social interaction and listening measures may also be explained, to some extent, by the skill acquisition theory of second language acquisition, which claimed language learning practice is skill-specific (DeKeyser, 2015). Social interaction, as a measure in our study, is a mixture of language input and output, while the specific weighting of these two components may vary across individuals. Empirical evidence suggests that input-based practice is more beneficial for receptive skills (i.e., listening and reading) and may have limited transfer to the improvement of productive skills (i.e., speaking, and writing), and vice versa (De Jong, 2005; DeKeyser & Sokalski, 2001; Rodgers, 2011). Therefore, input-based learning practice (e.g., listening to broadcast, watching TV shows, and attending to lectures) may be a more effective means for advancing listening skills than social interaction. Hence, this could, to some degree, be taken as evidence for the skill-specific effect of language learning practice.

Secondly, in the meta-analysis significant SA effects were found for L2 production but not for comprehension, although the moderation effect of language dimension just missed significance. It seems to suggest that compared to an AH learning context, the SA learning context tends to be more beneficial for the development of production skills than for that of comprehension skills. This finding thus seems to point in the direction of skill specificity, which claims that certain kinds of language-learning practice may not benefit language comprehension and production to the same extent (DeKeyser, 2007, 2015). Future studies are clearly needed to further investigate the possibility of skill-specific effects of language learning in the study-abroad learning contexts and beyond.

Thirdly, DeKeyser's claim about skill specificity has mainly been validated by comparing comprehension and production skills, whereas I am also interested in comparing the listening subprocesses (i.e., lexical access, grammatical processing, and semantic proposition formation). At pretest, the biggest difference in language processing between native and nonnative groups was found in speed and stability of word recognition, whereas grammar sensitivity was the subprocess that least distinguished the proficiency of nonnative groups (Chapter 2). However, we have to bear in mind that the three tasks (and also the vocabulary size test) were administered in a fixed order and they differed in measurement characteristics such that task difficulty may be confounded with order effects. Nevertheless, the nonnative groups did not differ significantly from each other in term of the pretest-to-posttest progress in these three listening subprocesses. This finding should be interpreted in the light of the fact that the pretest and posttest were the same and hence participants' progress in the processing tasks may be generally inflated due to the test-retest effect (elaborated on in section 6.3).

Apart from skill-based theories (Anderson, 1983; DeKeyser, 2015), usage-based theories of second language acquisition also concern the relationship between language exposure and second language development (Chapter 2). On the one hand, we found a pattern of decreasing vocabulary sizes from the AH-intensive group to the SA-onset group and then to the AH-regular group, which was likely due to their decreasing exposure levels conditioned by learning contexts. This finding suggests that more L2 exposure in EFL learning contexts translates into more vocabulary knowledge, which agrees with usage-based theories on the role of exposure in vocabulary acquisition (e.g., N. Ellis, 2006). This finding is also in line with previous studies on the correlation between exposure and reading vocabulary size (e.g., Milton, 2009). On the other hand, we found that the AH-intensive and SA-onset groups did not differ from each other in processing efficiency, but they both outperformed the AH-regular group. This finding suggests that more exposure in EFL learning contexts does not necessarily lead to higher language processing efficiency. Note that, since the performance of nonnative groups in our study was still quite distant from ceiling performance as suggested by native performance, there still seemed to be ample room for improvement. Hence, our results seem to partly contradict the usage-based accounts (e.g., Bybee, 2003, 2006; N. Ellis, 2006; Schmid, 2007) on the facilitative role of exposure in language processing. Possibly, however, in order to interpret these results, we should not only take quantity of exposure into account, but also type of exposure. Apart from amount of exposure, the target language exposure in different EFL learning contexts may also differ in terms of the characteristics (i.e., token and type frequencies) of words and grammatical constructions and in terms of the ratio of auditorily- and visually-presented input. Differences in type of exposure across learning contexts may also contribute to differential effects of learning contexts on vocabulary and processing efficiency.

Relatedly, the second way in which this thesis speaks to usage-based theories is that native and nonnative language processing differed the most in speed of word recognition (Chapter 2). Note that this result has to be interpreted with caution due to the fact that we administered the tasks in a fixed order and the lexical access task was administered first. Usage-based theories of second language acquisition emphasize the role of token frequency in lexical entrenchment (e.g., Bybee, 2006; Schmid, 2007). The failure of EFL learners to achieve more efficient word recognition might be associated with the low token frequency of lexical items in their input compared to naturalistic input. Future research is needed to investigate how to help EFL learners to entrench word representations and recognition.

6.3 Methodological concerns and limitations

It is methodologically challenging to investigate whether studying abroad is beneficial for a certain aspect of language learning. A common problem associated with study-abroad research is that conclusions on linguistic benefits of studying abroad were sometimes drawn with methodological designs that may be flawed to serve this specific research purpose. Two types of designs were found to carry the risk of inflating the observed study-abroad effect size (see Chapter 5): between-group posttest-only designs comparing SA and AH groups only once without taking into account baseline group differences (see also Chapter 2 and 3) and within-group pretest-posttest designs comparing the pretest and posttest performance of an SA group without including an AH comparison group. We employed a between-group pretest-posttest design to study the effect of studying abroad, thus minimizing research biases stemming from choice of methodological designs.

Another methodological issue of study-abroad research is that researchers have to deal with existing groups in different learning contexts and it is practically impossible to randomly assign participants to different groups. That is, lack of randomization is inherent to study-abroad research, which may lead to between-group differences in various aspects, e.g., baseline language proficiency, language aptitude, learning motivation, and personalities. Groups may very well differ in these aspects, especially in baseline language proficiency. Since study-abroad learners have to spend more time and effort learning the target language in preparation for studying abroad, study-abroad learners may already outperform their at-home counterparts at baseline, which was definitely the case in our studies. Although group differences in baseline language proficiency can be partly addressed by taking into account pretest performance (as done in the present study), they cannot be completely addressed because baseline proficiency may also interact with amount of progress made by participants. Therefore, though non-English-major study abroad learners are conventionally compared against non-English-major at-home learners in study-abroad research, we decided to have two at-home control groups whose baseline proficiency levels were expected to be above and below that of the study-abroad group respectively, in order to avoid either overestimating or underestimating the effect of studying abroad.

However, having two control groups is not the same as randomly assigning participants to groups. One possible way to get around this no-randomization problem is to conduct experimental studies where learning conditions can be manipulated, participants can be randomly assigned to groups, and third factors (e.g., baseline language proficiency) can be controlled for. We recommend future research to extend this line of research into investigating the effect of learning conditions (e.g., online educational interventions) on L2 listening development with carefully-controlled experimental designs. Nevertheless, note that experimentally-manipulated learning conditions may simulate study abroad settings but may suffer from ecological validity issues to some extent. Therefore, our recommendation is not to undermine the importance of study-abroad research, but to encourage researchers to synergize studies of natural learning contexts (e.g., study-abroad contexts) and those of experimental learning conditions.

Furthermore, the validity of measurement is of paramount importance for any empirical research. Firstly, in the present thesis, the absolute amount of progress participants made at posttest may be influenced by test-retest effects of being presented with the same materials twice (see Chapter 3). In future SA research, testing materials could be counterbalanced over time points and participants to minimize effects of test material familiarity. Relatedly, counterbalancing task order could avoid the potential confound of task order on performance differences among tasks, as discussed previously.

Secondly, zooming in on our studies, the three processing tasks, measuring spoken-language processing efficiency in lexical access, grammatical processing and semantic processing, deserve some discussion with respect to their ecological validity (i.e., whether the results can be applied to real-life listening comprehension). Participants responded to the testing stimuli by pressing corresponding buttons and had to memorize which button represented what (i.e., match or not, left or right, and plausible or not). In real-life listening people do not need to constantly make explicit decisions on what they hear by pressing buttons. Two out of the three processing tasks used picture-matching paradigms, whereas listening comprehension does not necessarily involve visual information processing. Therefore, there are other processes than language processing involved, which may undermine the ecological validity of these tasks. Future studies may want to consider using other methods like eye-tracking methods to measure 'real-life listening' (without language learners having to make explicit decisions) or to combine the picture-matching paradigm we used with more holistic measurement methods like spoken-passage comprehension.

Apart from the points about test-retest effects and ecological validity of testing instruments, the assumption that the coefficient of variation (CV) can be interpreted as a measure of processing stability also needs some further discussion. Assuming that automatic processing is characterized by reorganizing or bypassing of serial execution of component processes, Segalowitz and Segalowitz (1993) proposed that decreases in CV values can be used to distinguish automatization (or restructuring) from simple speeding up. However, empirical studies have not provided unequivocal evidence for automatization through analysis of CVs (see Hulstijn et al, 2009 for a review). The interpretation of CV as a measure of degree of automatization seems to have at least three limitations. Firstly, the suitability of using CV for measuring automatization in higher-level processes, e.g., employing background knowledge to draw inferences, is questionable (Lim & Godfroid, 2015). Lim and Godfroid argued against the possibility that higher-level processes can be automatized in the first place, because such processes require conscious processing and are highly context-dependent. Lim and Godfroid also demonstrated that CV may be an effective measure for lower-level processes (e.g., word recognition, parsing, and semantic proposition formation, as tested in our study). As such, one should consider that the CV measure may not apply to all language processes. Therefore, when designing testing materials to measure language processing, we focused on testing lower-level processes and tried to reduce the involvement of higher-level processes. Secondly, CV analyses may be particularly problematic when applied to low-proficiency learners, for whom "gains in knowledge itself and gains in processing it cannot be adequately disentangled" (Hulstijn et al., 2009). In other words, improvement of declarative knowledge here is confounded with measures of processing stability, which makes the CV measure less widely applicable. Thirdly, and possibly relatedly, in the present study, out of the three measures of processing efficiency (i.e., accuracy rate, reaction time, and CV), CV has shown to be the least effective in differentiating the participant groups from different learning contexts. We argue that the coefficient of variation may be a valid but not particularly sensitive measure of processing stability or automatization, as it can generally be used to differentiate the natives and nonnative groups, but not the nonnative groups whose proficiency levels are not widely different. Future researchers interested in CV analyses are advised to consider how appropriate the CV measure may be given the proficiency level of their participants and the involvement of higher-level processes in their testing materials, as well as to explore more sensitive measures of processing stability.

Finally, how to reliably measure language exposure is a general problem in study-abroad research. It is inherently difficult to quantify the exact difference in language exposure between the nonnative groups because measuring language exposure concerns not only the exposure during the specific research period but also participants' language learning history. In Chapter 2, where we assumed exposure differences between groups, there was no fine-grained measurement of participants' language exposure in each group. However, judging from their self-reported estimation of instructed and self-directed language exposure and from the result patterns of the vocabulary test, we do think our statement that the exposure level of the SA-onset group should be in between that of the other two groups is justified. Additionally, we interpret our results taking into account not only quantity of exposure but also type of exposure (see section 6.2). Nevertheless, future research may want to try to quantify L2 exposure in a more reliable way, e.g., by capturing everyday language exposure and use over a period of time with a digital app, as done in the currently (2021) ongoing LANG-TRACK-APP project (Arndt et al, in press).

6.4 Study abroad: What to expect?

Study abroad has been the central topic of this thesis. This final section will review what we have learned about study-abroad participants and the efficacy of studying abroad in facilitating second language development. The section will end with some concluding remarks.

As mentioned before, with regard to our study-abroad participants, we found that they already outperformed their at-home non-English-major peers in listening vocabulary size and spoken-language processing efficiency at the onset of their period abroad. This outperformance was likely due to their active preparation for studying abroad. Moreover, the processing efficiency of the study-abroad participants was already equal to that of domestic English majors, although their vocabulary sizes were still smaller than those of domestic English majors. This seems to suggest that language exposure in EFL (at-home) contexts has a limited effect on developing language processing skills for intermediate-to-advanced learners (e.g., the domestic English majors). Note that the Peabody Picture Vocabulary Test showed that the average vocabulary scores of the study-abroad group at pretest were similar to those of 8 year-old native speakers (more precisely: age equivalence was 8 years and 3 months).

During their one-year abroad, the SA students acquired both listening vocabulary and improved their spoken-language processing efficiency. They made more improvement in processing efficiency than the domestic English majors who had equal processing efficiency at the beginning of this study. However, their improvement in vocabulary size seemed small and not particularly larger than the domestic non-English-major group, with their native age equivalence at the posttest increasing by 6 months to 8 years and 9 months. This suggests that study abroad is an effective intervention for developing L2 processing efficiency but not necessarily for vocabulary acquisition. Meanwhile, in terms of individual variability, learners with higher language aptitude were associated with larger vocabulary sizes and higher processing efficiency, and learners with higher amount of L2 exposure were associated with higher processing efficiency. There seems to be no differences in the way these individual-difference factors were associated to L2 listening proficiency across learning contexts.

With regard to the efficacy of study-abroad programs on second language development, we found that learners indeed benefited from participating in these study-abroad programs. However, the linguistic benefits of studying abroad for a semester or a year may not be as great as commonly believed. This was backed up by the meta-analysis finding that the overall study-abroad effect was small-to-medium. It should be noted, however, that the duration of study-abroad programs mattered, with longer programs being more effective in facilitating L2 development than short-term ones.

Furthermore, study-abroad programs seemed to be more beneficial for developing general proficiency and processing aspects of language learning than for facilitating knowledge aspects of language learning. It is promising to have found that studying abroad is effective in facilitating L2 processing, an aspect that learners in EFL learning contexts often stumble on. At the same time, study-abroad learners should be aware that studying abroad may not be particularly helpful for accumulating knowledge of a target language. Study-abroad learners may therefore need to make conscious efforts if they wish to gain maximal benefits from their study-abroad experiences.

Finally, I would like to end this thesis by looking back at the “dirty-glass” problem I started this thesis with. In order to address this problem, I have adopted the theoretical frameworks of the skill-based and the usage-based accounts of second language acquisition. This dirty-glass problem, triggered by weak listening proficiency, is both a problem of lacking sufficient knowledge of the target language and lacking efficient language processing skills. Through empirical investigations and systematic review of previous literature, we have identified problems with foreign language (or at-home) learning contexts, benefits and limits of study-abroad learning contexts, as well as factors accounting for individual variability in second language attainment. It seems studying abroad for an academic year indeed helps to clean the “dirty glass” of second language listening, but only to a limited extent. Future studies may want to further investigate how to optimize learning contexts, e.g., through online education interventions, in order to alleviate language-learning problems associated with these existing learning contexts.

Nederlandse samenvatting

Studeren in het buitenland wordt vaak beschouwd als de beste context voor het leren van een tweede taal (T2), omdat leerlingen hun eerste taal moeten onderdrukken en zich moeten onderdompelen in de tweede taal (Freed, 1995; Jacobs, Fricke, & Kroll, 2016; Linck, Kroll, & Sunderman , 2009). Het leren van een andere taal in het thuisland oftewel in een vreemde-taalcontext kan worden bekritiseerd vanwege de relatief beperkte blootstelling aan de doeltaal, het te veel vertrouwen op uit het hoofd leren en de ontoereikende mogelijkheden voor interactie. Verschillen in taalleercontexten kunnen resulteren in verschillende kenmerken en trajecten van de tweedetaalontwikkeling (Kroll, Dussias, & Bajo, 2018), en kunnen zelfs van invloed zijn op hoe individuele verschillen (zoals in werkgeheugen) geassocieerd zijn met T2-leerprocessen en -uitkomsten (Faretta-Stutenberg & Morgan-Short, 2018; Sunderman & Kroll, 2009).

Dit proefschrift onderzoekt deze veronderstellingen over leercontext, individueel-verschilfactoren en T2-leren, met de nadruk op het leren van een taal in het betreffende buitenland. Hoofdstukken 2 tot en met 4 presenteren een empirisch onderzoek dat ingaat op luistervaardigheid, het minst onderzochte terrein op het gebied van tweedetaalverwervingsonderzoek. De weinige luisteronderzoeken die er zijn, hebben de T2-luistervaardigheid meestal op een holistische manier gemeten (bijvoorbeeld met opdrachten voor het begrijpen van gesproken passages). Het huidige onderzoek heeft luistervaardigheid geanalyseerd vanuit het perspectief van theorieën over de verwerving van cognitieve vaardigheden (Anderson, 1983; DeKeyser, 2015) en van theorieën over luistervaardigheid (Anderson, 2015; Goss, 1982; Cutler & Clifton, 1999). Meer specifiek werd de T2-luistervaardigheid in dit proefschrift geoperationaliseerd als een vorm van kennis , namelijk de luisterwoordenschat, en als verwerkingsefficiëntie van gesproken taal. In Hoofdstuk 5 hebben we uitgezoomd door een systematische literatuurreview uit te voeren van bestaand onderzoek naar de effecten van buitenlandverblijf ('study abroad') op tweedetaalontwikkeling. Hierna zal ik voor elke deelstudie een samenvatting geven van onze belangrijkste bevindingen.

Hoofdstuk 2 onderzocht of leerders uit drie Engels-als-vreemde-taal (EFL) leercontexten verschillen in hun kennis en verwerkingsefficiëntie van EFL luisterstimuli. Drie groepen Chinese masterstudenten hebben samen met een controlegroep van moedertaalsprekers van het Engels een batterij luistertests gedaan. De Chinese groepen omvatten niet-Engelse majors die in China studeerden ('at home', AH-reguliere groep), niet-Engelse majors die pas in het Verenigd Koninkrijk waren aangekomen ('study abroad', SA-onset-groep) en Engelse majors die in China studeerden (AH-intensieve groep). Op basis van het aantal uren instructie en zelfgestuurde leeractiviteiten gingen we ervan uit dat de drie EFL-leercontexten verschilden met betrekking tot de hoeveelheid T2-blootstelling: de AH-intensieve groep met de meeste blootstelling, de AH-reguliere groep met de minste blootstelling, en de SA-onset-groep, die ergens tussen deze twee AH-groepen ligt (vanwege hun voorbereiding op studeren in het buitenland). De luistertests maten auditieve woordenschatkennis en luisterverwerkingsefficiëntie (d.w.z. correctheid, snelheid en stabiliteit van T2-spraakverwerking) in lexicale toegang, grammaticale verwerking en semantische verwerking.

Er waren drie belangrijke bevindingen. Ten eerste bereikte geen van de drie EFL-groepen het niveau van de moedertaalgroep: noch in termen van woordenschatgrootte, noch in termen van correctheid, snelheid en stabiliteit van verwerking. Ten tweede bleek de AH-intensieve groep een grotere woordenschat te hebben dan de SA-onset-groep, die op zijn beurt beter presteerde dan de AH-reguliere groep. Ten derde verschilden de AH-intensieve en de SA-onset-groep in geen van de verwerkingsmaten, maar ze presteerden beide beter dan de AH-reguliere groep in correctheid en snelheid van verwerking in de drie verwerkingstaken. Deze resultaten lijken te suggereren dat meer EFL-blootstelling gerelateerd is aan een grotere woordenschat, maar niet noodzakelijk aan een grotere verwerkingsefficiëntie van gesproken taal. Al met al verschilden kennis- en verwerkingsaspecten van T2-luistervaardigheid tussen groepen en worden ze mogelijk niet in gelijke mate beïnvloed door EFL-blootstelling, wat een beperkt effect lijkt te hebben op het faciliteren van taalverwerkingsvaardigheden voor halfgevorderde tot gevorderde leerders.

Hoofdstuk 3 onderzocht het effect van studeren in het buitenland op de T2-luisterontwikkeling. Dit hoofdstuk bouwde voort op het vorige hoofdstuk doordat de deelnemers in hoofdstuk 2 werden uitgenodigd om dezelfde testen na ongeveer een studiejaar opnieuw af te leggen. Dit stelde ons in staat om T2-luisterontwikkeling te vergelijken met betrekking tot taalkennis en verwerkingsefficiëntie in één studie-buitenlandcontext (SA) en twee thuisblijf contexten (AH) in de loop van een studiejaar. De SA-onset-groep in hoofdstuk 2 werd hernoemd tot de SA-groep in hoofdstuk 3.

Wat betreft woordenschatontwikkeling, vonden we dat de SA-groep meer vooruitgang boekte dan de AH-intensieve groep, maar niet significant verschilde van de AH-reguliere groep. De groei van de woordenschat kan echter interacteren met pre-testscores, wat in onze data moeilijk te onderscheiden is van het effect van leercontext. Desalniettemin lijken de resultaten te suggereren dat de voordelen van studeren in het buitenland voor de verbetering van de woordenschat klein zijn of niet bestaan. Met betrekking tot de verwerving van T2-verwerkingsefficiëntie ontdekten we dat de SA-groep meer vooruitgang boekte dan de AH-intensieve groep, maar minder vooruitgang dan de AH-reguliere groep, in snelheid van verwerking over taken heen. Aangezien de SA-groep en de AH-intensieve groep met gelijke verwerkingsefficiëntieniveaus tijdens de pre-test begonnen, betogen we dat hun verschillende ontwikkelingspatronen kunnen worden toegeschreven aan het effect van de leercontext. Aangezien steile leercurves voor deelnemers met lage taalverwerkingssnelheid vaak zijn waargenomen in experimentele settings, argumenteren we daarnaast dat de AH-reguliere groep meer verbetering boekte dan de SA-groep vanwege de lagere verwerkingssnelheid bij de pre-test. Echter, de drie groepen boekten gelijke vooruitgang in correctheid en stabiliteit van de verwerking. Deze resultaten suggereren dat studeren in het buitenland, bij gelijke startniveaus, gunstiger is voor het verbeteren van de T2-verwerkingssnelheid dan thuisblijven. Alles bij elkaar genomen toonde Hoofdstuk 3 aan dat studeren in het buitenland een effectieve methode kan zijn voor het verwerven van T2-verwerkingsefficiëntie (met name snelheid), maar niet noodzakelijk voor het verwerven van woordenschat.

Hoofdstuk 4 was gericht op het onderzoeken van de effecten van individuele verschillen in achtergrondvariabelen op T2-begrijpend luisteren in buitenlandstudie (SA) en thuisstudie (AH) leercontexten, waarbij de mogelijke wisselwerking tussen individuele verschillen en leercontext werd onderzocht. Dit hoofdstuk bouwde voort op het vorige hoofdstuk in die zin dat dezelfde deelnemers tijdens de post-test individuele-verschillentests en vragenlijsten moesten maken. Omdat T2-leerders taal met verschillende snelheden kunnen verwerven als gevolg van individuele verschillen in taalvaardigheid, werkgeheugen, mentaal welzijn, taalblootstelling en sociale interactie, hebben we getest hoe goed deze individuele verschillen de prestaties van de deelnemers in de pre- en post-test voorspelden en hun verbetering in verschillende leercontexten.

Ten eerste bleek aanleg positief te correleren met de luisterwoordenschat van de deelnemers en hun verwerkingsefficiëntie, om precies te zijn de correctheid in alle drie de luistertaken en de verwerkingssnelheid in twee van de drie taken. Dit suggereert dat aanleg niet alleen betrekking heeft op maten van gekristalliseerde taalvaardigheid (inclusief woordenschatgrootte), maar ook op de efficiëntie waarmee de T2 wordt verwerkt. Ten tweede bleek het werkgeheugen niet gerelateerd aan de luistermetingen. Dit komt waarschijnlijk doordat aanleg en werkgeheugen beide in de statistische modellen zijn opgenomen, waarbij de twee een aanzienlijke hoeveelheid gedeelde voorspelde variantie hebben (zie Mackey, Adams, Stafford, & Winke, 2010). Ten derde was taalblootstelling (gekwantificeerd als frequentie van het uitvoeren van taalleeractiviteiten) geassocieerd met verwerkingsefficiëntie. Deze associatie kan worden verklaard door het centrale concept van de theorie voor het verwerven van (tweedetaal)vaardigheden, waarin wordt gesteld dat de reactietijd en het foutenpercentage afnemen door te oefenen (DeKeyser, 2015). Ten vierde bleek de hoeveelheid sociale interactie niet significant gerelateerd aan de luistermetingen, wat suggereert dat T2-leerders met meer sociale interactie niet noodzakelijkerwijs een hogere luistervaardigheid hebben. Het nuleffect van sociale interactie moet met voorzichtigheid worden geïnterpreteerd, want deze variabele, geoperationaliseerd als de geschatte spreektijd van deelnemers met anderen, kan voor deelnemers moeilijk zelf te rapporteren zijn geweest. Ten vijfde was mentaal welbevinden ook niet gerelateerd aan de luistermetingen. Dit suggereert dat mentaal welbevinden ofwel niet relevant was voor het T2-leren van volwassenen of dat deelnemers hun persoonlijke gevoelens mogelijk niet in de vragenlijsten hebben onthuld. Tot slot vonden we, onverwacht, geen interactie tussen de effecten van individuele verschillen en leercontext, wat betekent dat leercontext geen invloed had op de manier waarop individuele verschillen verband hielden met luistervaardigheid of de ontwikkeling ervan. Samengevat liet Hoofdstuk 4 liet zien dat, hoewel geen van de individuele-verschilfactoren de ontwikkeling gedurende het academische jaar voorspelde, sommige (d.w.z. aanleg en blootstelling) stabiele voorspellers waren van luistervaardigheid bij zowel pre- als posttest in de leercontexten. Het leren van een tweede taal lijkt dus op vergelijkbare wijze verband te houden met de capaciteiten van individuen in verschillende leercontexten.

Naast bovengenoemde empirische onderzoeken heb ik ook een systematische review gedaan van eerdere literatuur over taalleereffecten tijdens studeren in het buitenland. **Hoofdstuk 5** nam bestaand onderzoek samen om een algemeen effect van studie in het buitenland op de T2-ontwikkeling te vinden (in vergelijking met het thuisbijven) en om factoren te identificeren die de omvang van de waargenomen effecten van studeren in het buitenland in individuele studies beïnvloedden. Twintig studies die de effectiviteit van studieprogramma's in het buitenland evalueerden bij het faciliteren van T2-ontwikkeling op universitair niveau, werden opgenomen in deze systematische review. Een multi-level meta-analyse leverde een klein tot middelgroot algemeen effect op van studeren in het buitenland op de ontwikkeling van een tweede taal. Dit suggereert dat studenten die in het buitenland studeren inderdaad baat hadden bij een onderdompeling in de doeltaal van de leeromgeving die wordt gekenmerkt door een hogere mate en kwaliteit van blootstelling aan de doeltaal, wat overeenkomt met de algemene opvatting over de voordelen van studeren in het buitenland voor het leren van talen. Bovendien bleek uit een aanvullende analyse dat het pretest-posttestonderzoeksontwerp met controlegroep de meest conservatieve schatting van de effecten van het buitenlandverblijf opleverde, gevolgd door het pretest-posttestonderzoeksontwerp binnen een groep, gevolgd door het tussen-groepenonderzoeksontwerp met alleen een post-test. Het pretest-posttestonderzoeksontwerp met controlegroep lijkt het meest geschikte onderzoeksontwerp, omdat het kan controleren voor algemene verbetering, ongeacht de leercontext, evenals voor verschillen tussen groepen vóór het vertrek. Moderatoranalyses toonden aan dat studies met langdurige SA-ervaringen (d.w.z. langer dan een semester) grotere effectgroottes lieten zien dan studies met korte SA-ervaringen (d.w.z. korter dan of gelijk aan één semester). Dit suggereert dat langetermijnverblijf in het buitenland effectiever is in het faciliteren van T2-ontwikkeling dan kort durende programma's. Bovendien lieten studenten die in het buitenland studeren, vergeleken met studenten thuis, meer vooruitgang zien op het gebied van algemene taalvaardigheid en maten gerelateerd aan taalverwerking dan op maten van kennis. Kortom, Hoofdstuk 5 rapporteerde een klein tot middelgroot algemeen effect van studeren in het buitenland op de ontwikkeling van een tweede taal, en identificeerde verschillende factoren (d.w.z. methodologisch ontwerp, verblijfsduur en type meting) die de grootte van de waargenomen SA-effecten medieerden.

Al met al levert dit proefschrift een bijdrage aan ons begrip van hoe leercontext en individuele capaciteiten samenhangen in het leren van een tweede taal. Door empirisch onderzoek te doen en een systematische review van eerdere literatuur uit te voeren, hebben we problemen geïdentificeerd met leercontexten voor vreemde-taalverwerving, voordelen en beperkingen van leercontexten in het buitenland, evenals factoren die verantwoordelijk zijn voor individuele variabiliteit in het verwerven van een tweede of vreemde taal. Toekomstige studies kunnen mogelijk verder onderzoeken hoe leercontexten kunnen worden geoptimaliseerd, bijvoorbeeld door online onderwijsinterventies, om taalleerproblemen die verband houden met deze bestaande leercontexten te verhelpen.

References

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*, *35*, 79–113.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford, England: Oxford University Press.

Anderson, J. R. (2015). *Cognitive psychology and its implications*. New York: Worth Publishers.

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, *62*(S2), 49–78.

Andringa, S., & Dąbrowska, E. (2019). Individual differences in first and second language ultimate attainment and their causes. *Language Learning, 69*(s2), 5–12.

Arndt, H., Granfelt, J., & Gullberg, M. (Accepted/In press). Reviewing the potential of the Experience Sampling Method (ESM) for capturing second language exposure and use. *Second Language Research.*

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Batty, A. O. (2021). Measure L2 listening. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 275–284). New York, NY: Routledge.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis.* Cornwall, UK: Wiley.

Brecht, R. D., & Robinson, J. L. (1995). On the value of formal instruction in study abroad. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 318–334). Amsterdam: John Benjamins.

Briggs, J. G. (2015). Out-of-class language contact and vocabulary gain in a study abroad context*. System, 53,* 129–140.

Buck, G. (2001). *Assessing listening*. Cambridge, England: Cambridge University Press.

Bybee, J. L. (2003). Mechanisms in change in grammaticalization: The role of repetition. In R. Janda & B. Joseph (Eds.), *Handbook of historical linguistics* (pp. 602–623). Oxford, England: Blackwell.

Bybee, J. L. (2006). *Frequency of use and the organization of language.* Oxford, England: Oxford University Press.

Bybee, J. L. (2013). Usage-based theory and exemplar representations. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 46–69). Oxford, England: Oxford University Press.

Bybee, J. L., & Hopper, P. J. (2001). *Frequency and the emergence of linguistic structure*. Philadelphia: John Benjamins.

Bybee, J., & Thompson, S. (2000). Three frequency effects in syntax. *Berkeley Linguistic Society*, *23*(1), 65–85.

Cable, D. M., Gino, F., & Staats, B. R. (2013). Breaking them in or eliciting their best? Reframing socialization around newcomers' authentic self-expression. *Administrative Science Quarterly, 58*(1), 1–36.

Carroll, J. B., & Sapon, S. M. (1959). *Modern Language Aptitude Test and Manual (MLAT)*. San Antonio, TX: The Psychological Corporation.

Carroll, J. B. (1990). Cognitive abilities in foreign language aptitude: Then and now. In T. Parry & C. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 11–29). Englewood Cliffs, NJ: Prentice Hall.

Carroll, S. (2001). *Input and Evidence: The raw material of second language acquisition*. Amsterdam: John Benjamins.

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. *Memory & Cognition*, *28*(1), 143–153.

Cho, K. S., & Krashen, S. D. (1994). Acquisition of vocabulary from the Sweet Valley Kids series: Adult ESL acquisition. *Journal of Reading, 37*(8), 662–667.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*(1), 3–42.

Cohen J. (1988). *Statistical power analysis for the behavioral sciences.* New York, NY: Routledge Academic.

Coleman, J. A. (2013). Researching whole people and whole lives. In C. Kinginger (Ed.), *Social and cultural aspects of language learning in study abroad* (pp. 17–44). Amsterdam: John Benjamins.

Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition, 26*(2), 227–248.

Collentine, J. (2009). Study abroad research: Findings, implications and future directions. In C. Doughty & M. Long (Eds.), *Handbook of language teaching* (pp. 218–233). Malden, MA: Blackwell.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Boston, MA: Houghton Mifflin.

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review*, *22*(5), 1216–1234.

Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*, *41*(4), 721–744.

Cubillos, J. H., Chieffo, L., & Fan, C. (2008). The impact of short-term study abroad programs on L2 listening comprehension skills. *Foreign Language Annals, 41*(1), 157–186.

Cubillos, J. H., & Ilvento, T. (2013). The impact of study abroad on students' self-efficacy perceptions. *Foreign Language Annals, 45*(4), 494-511.

Cutler, A., & Charles, C. (1999). Comprehending spoken language: a blueprint of the listener. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 123–166). Oxford, England: Oxford University Press.

Cutrone, P., & Datzman, B. R. (2015.). Japanese EFL university students and the study abroad experience: Examining L2 development and program satisfaction after three weeks in North America. *TESOL International Journal, 10*(2), 24–47.

Dahlen, K., & Caldwell–Harris, C. (2013). Rehearsal and aptitude in foreign vocabulary learning. *The Modern Language Journal*, *97*(4), 902–916.

Davidson, D. E. (2010). Study abroad: When, how long, and with what results? New data from the Russian front. *Foreign Language Annals, 43*(1), 6–26.

Dąbrowska, E. (2019). Experience, aptitude, and individual differences in linguistic attainment: A comparison of native and nonnative speakers: experience, aptitude, and individual differences. *Language Learning*, *69*(S1), 72–100.

Dąbrowska, E., & Street, J. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and nonnative English speakers. *Language Sciences*, *28*(6), 604–615.

De Jong, N. (2005). Can second language grammar be learned through listening? An experimental study. *Studies in Second Language Acquisition, 27*(2), 205–234.

de la Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in the receptive and productive acquisition of words. *Studies in Second Language Acquisition, 24*(1), 81–112.

DeKeyser, R. M. (1991). The semester overseas: What difference does it make? *ADFL Bulletin, 22*(2), 42–48.

DeKeyser, R. M. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition, 19*(2), 195–221.

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning*, *55*(S1), 1–25.

DeKeyser, R. M. (Ed.). (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge, England: Cambridge University Press.

DeKeyser, R. (2010). Monitoring processes in Spanish as a second language during a study abroad program. *Foreign Language Annals, 43*(1), 80-92.

DeKeyser, R. M. (2012). Interactions between individual differences, treatments, and structures in SLA. *Language Learning, 62*(s2), 189-200.

DeKeyser, R. M. (2014). Research on language development during study abroad: Methodological consideration and future perspectives. In C. Pérez-Vidal (Ed.), *Language acquisition in study abroad and formal instruction contexts* (pp. 313–325). Amsterdam: John Benjamins.

DeKeyser, R. M. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 94–112). New York, NY: Routledge.

DeKeyser, R., & Koeth, J. (2011). Cognitive aptitudes for second language learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 395–406). New York, NY: Routledge.

DeKeyser, R. M., & Sokalski, K. J. (2001). The differential role of comprehension and production practice. *Language Learning, 51*, 81–112.

Dewaele, J. -M. (2002). Psychological and sociodemographic correlates of communicative anxiety in L2 and L3 production. *International Journal of Bilingualism, 6*(1), 23–38.

Dewaele, J. -M. (2009). Individual differences in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 623–646). Bingley, UK: Emerald Group Publishing Limited.

Dewaele, J. -M., Chen, X., Padilla, A. M., & Lake, J. (2019). The Flowering of Positive Psychology in Foreign Language Teaching and Acquisition Research. Frontiers in Psychology, 10, Article 2128.

Dewey, D. P. (2004). A comparison of reading development by learners of Japanese in intensive domestic immersion and study abroad contexts. *Studies in Second Language Acquisition, 26*(02), 303–327.

Dewey, D. P. (2008). Japanese vocabulary acquisition by learners in three contexts. Frontiers: *The Interdisciplinary Journal of Study Abroad, 15,* 127–148.

Dewey, D. P., Belnap, R. K., & Hillstrom, R. (2013). Social network development, language use, and language acquisition during study abroad: Arabic language learners' perspectives. *Frontiers: The Interdisciplinary Journal of Study Abroad, 22*(1), 84–110.

Diao, W., Freed, B., & Smith, L. (2011). Confirmed beliefs or false assumptions? a study of home stay experiences in the French study abroad context. *Frontiers: The Interdisciplinary Journal of Study Abroad, 21*(1), 109–142.

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, *66*(5), 843–863.

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2019). Prediction and integration of semantics during L2 and L1 listening. *Language, Cognition and Neuroscience, 34*(7), 881–900.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition.* Mahwah, NJ: Erlbaum.

Dörnyei, Z. (2009). Individual differences: Interplay of learner characteristics and learning environment. *Language learning, 59*(s1), 230-248.

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test.* Minneapolis, MN: Pearson Assessments.

Dupuy, B., & Krashen, S. D. (1993). Incidental vocabulary acquisition in French as a foreign language. *Applied Language Learning, 4*, 55–63.

Dwyer, M. (2004). More is better: The impact of study abroad program duration. *Frontiers: The Interdisciplinary Journal of Study Abroad, 10,* 151–163.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual Learning. *Applied Linguistics*, *27*(2), 164–194.

Ellis, R. (2016). Focus on form: A critical review. *Language Teaching Research*, *20*(3), 405–428.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). Second language verb-argument constructions are sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism*, *4*(4), 405–431.

Ellis, R., Tanaka, Y., & Yamazaki, A. (1994). Classroom interaction, comprehension, and the acquisition of L2 word meanings. *Language Learning, 44*(3), 449–491.

Eurostat: Statistics explained. (2020, October). *Learning mobility statistics.* https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Learning_mobility_statistics#:~:text=1.3%20million%20students%20from%20abroad,the%20EU%2D27%20in%202018.&text=In%202018%2C%2023%20%25%20(312,in%20Italy%20and%20the%20Netherlands.

Faretta-Stutenberg, M., & Morgan-Short, K. (2018). The interplay of individual differences and context of learning in behavioral and neurocognitive second language development. *Second Language Research*, *34*(1), 67–101.

Farrell, T. S., & Mallard, C. (2006). The use of reception strategies by learners of French as a foreign language. *The Modern Language Journal, 90*(3), 338–352.

Farshi, N., & Tavakoli, M. (2019). Effects of differences in language aptitude on learning grammatical collocations under elaborated input conditions. *Language Teaching Research, 25*(3), 476–499.

Fenson L., Marchman, V. A. , Thal, D., Dale, P. S., Reznick, J. S., Bates, E. (2006). *MacArthur-Bates communicative development inventories: User's guide and technical manual*. Baltimore, MD: Brookes Publishing Co.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234–248.

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, *42*(1), 98–116.

Félix-Brasdefer, J. C., & Hasler-Barker, M. (2015). Complimenting in Spanish in a short-term study abroad context. *System, 48*, 75–85.

Foster, P., Bolibaugh, C., & Kotula, A. (2014). Knowledge of nativelike selections in a L2: The influence of exposure, memory, age of onset, and motivation in foreign language and immersion settings. *Studies in Second Language Acquisition, 36*(1), 101–132.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist, 56*(3), 218.

Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & Emotion, 19*(3), 313–332.

Freed, B. F. (Ed.). (1995). *Second language acquisition in a study abroad context*. Philadelphia/Amsterdam: John Benjamins.

Freed, B. F. (1998). An overview of issues and research in language learning in a study abroad setting. Frontiers: The interdisciplinary journal of study abroad, 4(1), 31–60.

Freed, B. F., Dewey, D. P., Segalowitz, N., & Halter, R. (2004). The language contact profile. *Studies in Second Language Acquisition, 26*(2), 349–356.

Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition, 26*(02), 275–301.

Freed, B., So, S., & Lazar, N. A. (2003). Language learning abroad: How do gains in written fluency compare with gains in oral fluency in French as a second language? *ADFL Bulletin, 34*(3), 34–40.

Garnier, M., & Schmitt, N. (2016). Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System*, *59*, 29–44.

Gass, S. (2002). Interactionist perspectives in SLA. In R. Kaplan (Ed.), *Handbook of applied linguistics* (pp. 170–181). Oxford, England: Oxford University Press.

Gass, S., & Mackey, A. (2007). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 180–206). New York: Routledge.

Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, *28*, 55–75.

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always a means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, *58*(3), 787–814.

Goss, B. (1982). Listening as information processing. *Communication Quarterly*, *30*(4), 304–307.

Granena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA Language Aptitude Test. In G. Granena & M. H. Long (eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 105–129). Amsterdam: John Benjamins.

Granena, G. (2016). Cognitive aptitudes for implicit and explicit learning and information-processing styles: An individual differences study. *Applied psycholinguistics, 37*(3), 577-600.

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research, 29*(3), 311–343.

Hahne, A. (2001). What's different in second-language processing? Evidence from event-related brain potentials. *Journal of Psycholinguistic Research*, *30*(3), 251–266.

Han, Z. (2013). Forty years later: Updating the Fossilization Hypothesis. *Language Teaching*, *46*(2), 133–171.

Han, Z., Park, E. S., & Combs, C. (2008). Textual enhancement of input: Issues and possibilities. *Applied Linguistics*, *29*(4), 597–618.

Han, Z., & Odlin, T. (Eds.). (2006). *Studies of fossilization in second language acquisition*. Clevedon, UK: Multilingual Matters.

Håkansson, G., & Norrby, C. (2010). Environmental influence on language acquisition: Comparing second and foreign language acquisition of Swedish. *Language Learning, 60*(3), 628–650.

Higby, E., & Obler, L. K. (2016). Length of residenceg. *Linguistic Approaches to Bilingualism, 6,* 43-63.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, England: John Wiley & Sons.

Hirakawa, M., Shibuya, M., & Endo, M. (2019). Explicit instruction, input flood or study abroad: Which helps Japanese learners of English acquire adjective ordering? *Language Teaching Research, 23*(2), 158–178.

Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations. London: Sage Publications.

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *The Canadian Modern Language Review, 61*(3), 355–382.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*(2), 207–223.

Howard, M. (2001). The effects of study abroad on the L2 learner's structural skills: Evidence from advanced learners of French. *EUROSLA Yearbook, 1*(1), 123–141.

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Malden, MA: Blackwell.

Hulstijn, J. H., van Gelderen, A., & Schoonen, R. (2009). Automatization in second-language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*(4), 555–582.

Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in Spanish-learning children. *Developmental Science*, *11*(6), F31–F39.

Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in SLA. In C. J. Doughty & M. H. Long (Eds), *The handbook of second language acquisition* (pp. 539–588). Oxford, England: Blackwell Publishing.

Ife, A., Vives Boix, G., & Meara, P. (2000). The impact of study abroad on vocabulary development among different proficiency groups. *Spanish Applied Linguistics, 4*(1), 55–84.

Isabelli-García, C. (2010). Acquisition of Spanish gender agreement in two learning contexts: Study abroad and at home. *Foreign Language Annals, 43*(2), 289–303.

Jacobs, A., Fricke, M., & Kroll, J. F. (2016). Cross-language activation begins during speech planning and extends into second language speech. *Language learning, 66*(2), 324–353.

Janacsek, K., & Nemeth, D. (2013). Implicit sequence learning and working memory: Correlated or complicated? *Cortex, 49*(8), 2001–2006.

Jiang, N. (2018). *Second language processing: An introduction*. New Yorker: Routledge.

Jochum, C. J. (2014). Measuring the effects of a semester abroad on students' oral proficiency gains: A comparison of at home and study abroad. *Frontiers: The Interdisciplinary Journal of Study Abroad, 24*(1), 93–104.

Kemmer, S., & Barlow, M. (2000). Introduction: A usage-based conception of language. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language*. Stanford, CA: CSLI.

Kersten, K. (2010). *ELIAS: Early language and intercultural acquisition studies. Final report: Public part.* Magdeburg, Germany: ELIAS.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences, 22*(2), 154–169.

Kinginger, C. (2009). *Language learning and study abroad: A critical reading of research*. New YorkL Palgrave Macmillan.

Kinginger, C. (2017). Language socialization in study abroad. In P. A. Duff & S. May (Eds.), *Language socialization* (pp. 1–12). Springer International Publishing.

Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing, 28,* 39–52.

Kogut, B., & Singh, H. (1988). The effect of national culture on the choice of entry mode. *Journal of International Business Studies, 19*(3), 411–432.

Krashen, S. (1985). *The input hypothesis: Issues and implications.* New York: Longman.

Kroll, J. F., Dussias, P. E., & Bajo, M. T. (2018). Language use across international contexts: Shaping the minds of L2 speakers. *Annual Review of Applied Linguistics, 38,* 60–79.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786.

Lafford, B. A. (2004). The effect of the context of learning on the use of communication strategies by learners of Spanish as a second language. *Studies in Second Language Acquisition, 26*(2), 201–225.

Lake, J. (2016). Accentuate the positive: Conceptual and empirical development of the positive L2 self and its relationship to L2 proficiency. In P. D. MacIntyre, T. Gregersen, & S. Mercer (Eds.), *Positive psychology in SLA* (pp. 237–257). Bristol, UK: Multilingual Matters.

Langacker, R. W. (1988). A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in cognitive linguistics*. Philadelphia: John Benjamins.

Langacker, R. W. (2000). A dynamic usage-based model. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language*. Stanford, CA: CSLI.

Lara, R., Mora, J. C. & Pérez-Vidal, C., (2015). How long is long enough? L2 English development through study abroad programmes varying in duration. *Innovation in Language Learning and Teaching, 9*(1), 46–57.

Lee, S. (2007). Effects of textual enhancement and topic familiarity on Korean EFL students' reading comprehension and learning of passive form. *Language Learning*, *57*(1), 87–118.

Legacy, J., Zesiger, P., Friend, M., & Poulin-Dubois, D. (2016). Vocabulary size and speed of word recognition in very young French–English bilinguals: A longitudinal study. *Bilingualism: Language and Cognition*, *21*(1), 137–149.

Leong, C.-H. (2007). Predictive validity of the multicultural personality questionnaire: A longitudinal study on the socio-psychological adaptation of Asian undergraduates who took part in a study-abroad program. *International Journal of Intercultural Relations, 31*(5), 545–559.

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition, 38*(4), 801–842.

Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, *36*(5), 1247–1282.

Linck, J. A., Kroll, J. F., & Sunderman, G. (2009). Losing access to the native language while immersed in a second language: Evidence for the role of inhibition in second-language learning. *Psychological science, 20*(12), 1507–1515.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21(4),* 861–883.

Linck, J. A., & Weiss, D. J. (2015). Can working memory and inhibitory control predict second language learning in the classroom? *Sage Open, 5*(4), 1–11.

Llanes, À. (2011). The many faces of study abroad: An update on the research on L2 gains emerged during a study abroad experience. *International Journal of Multilingualism, 8*(3), 189-215.

Llanes, À., & Muñoz, C. (2009). A short stay abroad: Does it make a difference? *System, 37*(3), 353–365.

Llanes, À., & Muñoz, C. (2013). Age effects in a study abroad context: Children and adults studying abroad and at home. *Language Learning, 63*(1), 63–90.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). San Diego, CA: Academic Press.

Lundell, F. F., & Lindqvist, C. (2014). Lexical aspects of very advanced L2 French. *Canadian Modern Language Review, 70*(1), 28-49.

Ma, D., Yu, X., & Zhang, H. (2017). Word-level and sentence-level automaticity in English as a foreign language (EFL) learners: A comparative study. *Journal of Psycholinguistic Research, 46*(6), 1471–1483.

MacIntyre, P. D. (2002). Motivation, anxiety and emotion in second language acquisition. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 45-68). Philadelphia: John Benjamins.

MacIntyre, P. D., Baker, S. C., Clement, R., & Donovan, L. A. (2003). Sex and age effects on willingness to communicate, anxiety, perceived competence, and L2 motivation among junior high school French immersion students. *Language Learning, 53*(S1), 137–165.

MacIntyre, P. D., Gregersen, T. (2012) Emotions that facilitate language learning: The positive-broadening power of the imagination. *Studies in Second Language Learning and Teaching, II*(2), 193–213.

MacIntyre, P. D., Gregersen, T., & Mercer, S. (2019). Setting an Agenda for Positive Psychology in SLA: Theory, Practice, and Research. *The Modern Language Journal, 103*(1), 262–274.

Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the Relationship Between Modified Output and Working Memory Capacity. *Language Learning, 60*(3), 501–533.

Maeder-Qian, J. (2018). Intercultural experiences and cultural identity reconstruction of multilingual Chinese international students in Germany. *Journal of Multilingual and Multicultural Development, 39*(7), 576–589.

Maertz, C. P., Hassan, A., & Magnusson, P. (2009). When learning is not enough: A process model of expatriate adjustment as cultural cognitive dissonance reduction. *Organizational Behavior and Human Decision Processes, 108*(1), 66–78.

Marchman, V. A., Fernald, A., & Hurtado, N. (2010). How vocabulary size in two languages relates to efficiency in spoken word recognition by young Spanish-English bilinguals. *Journal of Child Language*, *37*(4), 817–840.

Marqués-Pascual, L. (2011). Study abroad, previous language experience, and Spanish l2 development. *Foreign Language Annals, 44*(3), 565–582.

Martinsen, R. A., Alvord, S. M., & Tanner, J. (2014). Perceived foreign accent: Extended stays abroad, level of instruction, and motivation. *Foreign Language Annals, 47*(1), 66–78.

Matsumura, S. (2001). Learning the rules for offering advice: A quantitative approach to second language socialization. *Language Learning, 51*(4), 635–679.

Matsuoka, W., & Hirsh, D. (2010). Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a foreign language, 22*(1), 56–70.

Mayer, R. (2009). *Multimedia Learning.* Cambridge, England: Cambridge University Press.

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language, 55*(3), 381–401.

McDonough, K., & Kim, Y. (2009). Syntactic priming, type frequency, and EFL learners' production of wh-questions. *The Modern Language Journal, 93*(3), 386–398.

McManus, K., Mitchell, R., & Tracy-Ventura, N. (2014). Understanding insertion and integration in a study abroad context: The case of English-speaking sojourners in France. *Revue Française de Linguistique Appliquée, XIX*(2), 97–116.

Meara, P. (2005). LLAMA language aptitude tests: The manual. Swansea, UK: Lognostics.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL - International Journal of Applied Linguistics, 107*(1), 17–34.

Mitchell, R., Tracy-Ventura, N., & McManus, K. (2017). *Anglophone Students Abroad: Identity, Social Relationships, and Language Learning.* Taylor & Francis.

Montero, L., Serrano, R., & Llanes, À. (2017). The influence of learning context and age on the use of L2 communication strategies. *The Language Learning Journal, 45*(1), 117–132.

Montrul, S., Foote, R., & Perpiñán, S. (2008). Gender agreement in adult second language learners and Spanish heritage speakers: The effects of age and context of acquisition. *Language Learning*, *58*(3), 503–553.

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly, 46*(4), 610–641.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364–386.

Muñoz, C., & Llanes, À. (2014). Study abroad and changes in degree of foreign accent in children and adults. The Modern Language Journal, 98(1), 432–449.

Muranoi, H. (2000). Focus on form through interaction enhancement: Integrating formal instruction into a communicative task in EFL classrooms. *Language Learning*, *50*(4), 617–673.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13.

Newton, J. (2013). Incidental vocabulary learning in classroom communication tasks. *Language Teaching Research, 17*(2). 164–187.

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do Things Fall Apart? *Reading in a Foreign Language, 22*(1), 31–55.

Peter, M. S., Durrant, S., Jessop, A., Bidgood, A., Pine, J. M., & Rowland, C. F. (2019). Does speed of processing or vocabulary size predict later language growth in toddlers? *Cognitive Psychology, 115,* Article 101238.

Pérez-Vidal, C. (2017). Study abroad and ISLA. In S. Loewen, & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 339-360). London: Routledge.

Pliatsikas, C., & Marinis, T. (2013a). Processing of regular and irregular past tense morphology in highly proficient second language learners of English: A self-paced reading study. *Applied Psycholinguistics, 34*(5), 943–970.

Pliatsikas, C., & Marinis, T. (2013b). Processing empty categories in a second language: When naturalistic exposure fills the (intermediate) gap. *Bilingualism: Language and Cognition, 16*(1), 167–182.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning, 61*(4), 993–1038.

Pütz, M., & Sicola, L. (Eds.). (2010). *Cognitive processing in second language acquisition: Inside the learner's mind*. Philadelphia/Amsterdam: John Benjamins.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reber, A., Walkenfield, F., & Hernstadt, R. (1991). Implicit and explicit learning: individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory and Cognition 17*(5), 888–96.

Rees, J., & Klapper, J. (2007). Analyzing and evaluating the linguistic benefit of residence abroad for UK foreign language students. *Assessment and Evaluation in Higher Education, 32*(3), 331–353.

Rees, J., & Klapper, J. (2008). Issues in the quantitative longitudinal measurement of second language progress in the study abroad context. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 89–105). London: Routledge.

Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition*, *19*, 223–247.

Robinson, P. (2001). Individual differences, cognitive abilities, aptitude complexes, and learning conditions in SLA. *Second Language Research, 17*(4), 368–392.

Robinson, P. (2002). Individual differences in intelligence, aptitude and working memory during adult incidental second language learning: A replication and extension of Reber, Walkenfeld,and Hernstadt (1991). In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 211–266). Amsterdam: Benjamins.

Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty & M. H. Long (Eds), *The handbook of second language acquisition* (pp. 539–588). Oxford, England: Blackwell Publishing.

Robinson, P. (2005). Aptitude and second language acquisition. *Annual Review of Applied Linguistics*, *25*, 46–73.

Robinson, P. (2013). Aptitude in second language acquisition. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 198–201). Malden, MA: Wiley-Blackwell.

Rodgers, D. M. (2011). The automatization of verbal morphology in instructed second language acquisition. *IRAL - International Review of Applied Linguistics in Language Teaching*, *49*, 295–319.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis. Publication bias in meta-analysis: Prevention, assessment and adjustments*. Hoboken, NJ: Wiley.

Ruan, J., & Leung, C. (2012). *Perspectives on teaching and learning English literacy in China*. New York, NY: Springer.

Sadoski, M. (2005). A dual coding view of vocabulary learning. *Reading & Writing Quarterly, 21*(3), 221–238.

Sagarra, N. (2008). Working memory and L2 processing of redundant grammatical forms. In: Z. Han (Ed.) *Understanding second language process* (pp. 133–47). Clevedon: Multilingual Matters.

Saito, K. (2015). Experience effects on the development of late second language learners' oral proficiency. *Language Learning, 65*(3), 563-595.

Sanz, C. (2005). Adult SLA: The interaction between external and internal factors. In C. Sanz (Ed.), *Mind and context in adult second language acquisition: Methods, theory, and practice* (pp. 3–20). Washington, DC: Georgetown University Press.

Sasaki, M. (2007). Effects of study-abroad experiences on EFL writers: A multiple-data analysis. *The Modern Language Journal, 91*(4), 602–620.

Sasaki, M. (2011). Effects of varying lengths of study-abroad experiences on Japanese EFL students' L2 writing ability and motivation: A longitudinal study. *TESOL Quarterly, 45*(1), 81–105.

Schmid, H. (2007). Entrenchment, salience and basic levels. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford handbook of cognitive linguistics* (pp. 117–138). Oxford, England: Oxford University Press.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329–363.

Segalowitz, N. (2003). Automaticity and second languages. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 382–408). Malden, MA: Blackwell.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge.

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition, 26*(2), 173–199.

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics, 14*(3), 369–385.

Selinker, L. (1972). Interlanguage. *IRAL - International Review of Applied Linguistics in Language Teaching*, *10*, 209–232.

Serrano, S. L. (2010). Learning Languages in Study Abroad And at Home Contexts: A Critical Review of Comparative Studies. *PORTA LINGUARUM 13*, 149–163.

Serrano, R., Llanes, À., & Tragant, E. (2011). Analyzing the effect of context of second language learning: Domestic intensive and semi-intensive courses vs. study abroad in Europe. *System, 39*(2), 133–143.

Sheen, Y. (2007). The Effect of Focused Written Corrective Feedback and Language Aptitude on ESL Learners' Acquisition of Articles. *TESOL Quarterly, 41*(2), 255–283.

Shenkar, O. (2012). Cultural distance revisited: Towards a more rigorous conceptualization and measurement of cultural differences. *Journal of International Business Studies, 43*(1), 1–11.

Shi, L. F., & Koenig, L. L. (2016). Acoustic-phonetic versus lexical processing in nonnative listeners differing in their dominant language. *American Journal of Audiology, 25*(3), 167–176.

Shiffrin, R. M., & Schneider, W. (1984). Automatic and controlled processing revisited. *Psychological Review*, *91*, 269–276.

Skehan, P. (2012). Language aptitude. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 381–395). London, UK: Routledge.

Skehan, P. (2014). *Individual differences in second language learning*. Routledge.

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: acquisition of collocations under different input conditions. *Language Learning*, *63*(1), 121–159.

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychon Bull Rev, 24*(4), 1077–1096.

Soylu, F. (2010) Forward / Backward Digit-Span Task (New Version). *Archives of Neurobehavioral Experiments and Stimuli,* 222.

Statista. (2020, October). *Number of students from China going abroad for study from 2008 to 2018.* https://www.statista.com/statistics/227240/number-of-chinese-students-that-study-abroad/#:~:text=In%202018%2C%20around%20662%2C100%20Chinese,international%20students%20in%20the%20world.

Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System, 93,* Article 102279.

Sunderman, G., & Kroll, J. F. (2009). When study abroad experience fails to deliver: The internal resources threshold effect. *Applied Psycholinguistics, 30*(1), 79–99.

Suzuki, Y., & DeKeyser, R. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning, 67*(4), 747–790.

Svalberg, A. M. L. (2007). Language awareness and language learning. *Language Teaching, 40*(4), 287–308.

Swanborn, M. S., & De Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research, 69*(3), 261–285.

Taatgen, N. A., & Lee, F. J. (2003). Production compilation: A simple mechanism to model complex skill acquisition. *Human Factors, 45*(1), 61–76.

Tagarelli, K. M., Ruiz, S., Moreno, J. L., & Rebuschat, P. (2016). Variability in second language learning: The roles of individual differences, learning conditions, and linguistic complexity. *Studies in Second Language Acquisition, 38*(Specia), 293–316.

Taguchi, N. (2011). The effect of L2 proficiency and study-abroad experience on pragmatic comprehension. *Language Learning, 61*(3), 904–939.

Taguchi, N., Xiao, F., & Li, S. (2016). Effects of intercultural competence and social contact on speech act production in a Chinese study abroad context. *The Modern Language Journal, 100*(4), 775–796.

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J., & Stewart-Brown, S. (2007). The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): Development and UK validation. *Health and Quality of Life Outcomes, 5*(1), 63.

Tokowicz, N., Michael, E. B., & Kroll, J. F. (2004). The roles of study-abroad experience and working-memory capacity in the types of errors made during translation. *Bilingualism: Language and Cognition, 7*(3), 255–272.

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*, 61–82.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Tomasello, M. (2009). The usage-based theory of language acquisition. In E. L. Bavin (Ed.), *The Cambridge handbook of child language* (pp. 69–87). Cambridge, England: Cambridge University Press.

Toomer, M., & Elgort, I. (2019). The development of implicit and explicit knowledge of collocations: A conceptual replication and extension of Sonbul and Schmitt (2013). *Language Learning*, *69*(2), 405–439.

Tseng, W.-T., Liu, Y.-T., Hsu, Y.-T., & Chu, H.-C. (2021). Revisiting the effectiveness of study abroad language programs: A multi-level meta-analysis. *Language Teaching Research*, Article 1362168820988423.

Tullock, B., & Ortega, L. (2017). Fluency and multilingualism in study abroad: Lessons from a scoping review. *System*, *71*, 7–21.

UK Parliament. (2021, February 15). *International and EU students in higher education in the UK FAQs.* https://commonslibrary.parliament.uk/research-briefings/cbp-7976/

van den Bosch, L. J., Segers, E., & Verhoeven, L. (2019). The role of linguistic diversity in the prediction of early reading comprehension: A quantile regression approach. *Scientific Studies of Reading, 23*(3), 203–219.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*, 191–210.

Varela, O. E. (2017). Learning outcomes of study-abroad programs: A meta-analysis. *Academy of Management Learning & Education, 16*(4), 531–561.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.

Viechtbauer. W. (2017, March 19). Morris (2008). The metafor Package: A Meta-Analysis Package for R. http://www.metafor-project.org/doku.php/analyses:morris2008

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130–163.

Waters, G. S., Caplan, D., & Rochon, E. (1995). Processing capacity and sentence comprehension in patients with alzheimer's disease. *Cognitive Neuropsychology*, *12*(1), 1–30.

Weber, A., & Broersma, M. (2012). Spoken word recognition in second language acquisition. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics.* Bognor Regis: Wiley–Blackwell.

Weber, A., & Cutler, A. (2004). Lexical competition in nonnative spoken-word recognition. *Journal of Memory and Language*, *50*(1), 1–25.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(3), 387–401.

Weist, R. M. (2002). Space and time in first and second language acquisition: A tribute to Henning Wode. In P. Burmeister, T. Piske, & A. Rohde (Eds.), *The integrated view of language development: Papers in honor of Henning Wode* (pp. 79–109). Trier: WVT Wissenschaftlicher Verlag Trier.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*(2), 250–271.

West, M. (1953). *A general service list of English words.* Longman, Green: Longman.

Williams, J. N. (2012). Working memory and SLA. In S. Gass & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 427–441). New York: Routledge.

Williams, T. R. (2005). Exploring the impact of study abroad on students' intercultural communication skills: Adaptability and sensitivity. *Journal of Studies in International Education, 9*(4), 356–371.

Winke, P. (2013). An investigation into second language aptitude for advanced Chinese learning. *The Modern Language Journal, 97*(1), 109–130.

Xu, Y. (2019). Changes in interlanguage complexity during study abroad: A meta-analysis. *System*, *80*, 199–211.

Yang, J.-S. (2016). The effectiveness of study-abroad on second language learning: A meta-analysis. *Canadian Modern Language Review*, *72*(1), 66–94.

Yi, W. (2018). Statistical sensitivity, cognitive aptitudes, and processing of collocations. *Studies in Second Language Acquisition, 40*(4), 831–856.

Appendix A (Chapter 2)

Table A1. Fixed-effect estimates of accuracy performance of native and nonnative groups in the three processing tasks

	β	***SE***	***z***	***p***
(Intercept)	3.46	0.13	26.52	< .001
GroupNonnativevsNative	-1.54	0.22	-7.17	< .001
Task2vs1	-0.57	0.25	-2.25	.024
Task3vs1	0.32	0.28	1.14	.256
GroupNonnativevsNative:Task2vs1	0.13	0.36	0.37	.712
GroupNonnativevsNative:Task3vs1	-0.47	0.44	-1.08	.280

*Note: Model specification in glmer(Accuracy ~ Group*Task + (1 + Group|Item_number) + (1 + Task|SubjectNo)).*

Table A2. Fixed-effect estimates of RT performance of native and nonnative groups in the three processing tasks

	β	***SE***	***t***	***p***
(Intercept)	3.20	0.39	8.19	< .001
GroupNonnativevsNative	0.34	0.03	10.47	< .001
Task2vs1	0.38	0.07	5.42	< .001
Task3vs1	0.31	0.08	4.10	< .001
log_audio_duration	0.61	0.06	11.12	< .001
Trial_number	-0.00	0.00	-8.55	< .001
GroupNonnativevsNative:Task2vs1	-0.13	0.04	-3.19	.002
GroupNonnativevsNative:Task3vs1	-0.17	0.04	-4.04	< .001

*Note: Model specification in lmer(log_RT ~ Group*Task + log_audio_duration + Trial_number + (1 + Task|SubjectNo) + (1+Group|Item_number)).*

Table A3. Fixed-effect estimates of CV performance of native and nonnative groups in the three processing tasks

	β	***SE***	***t***	***p***
(Intercept)	0.30	0.01	59.02	< .001
GroupNonnativevsNative	0.06	0.01	6.15	< .001
Task2vs1	0.09	0.01	9.61	< .001
Task3vs1	-0.05	0.01	-5.17	< .001
GroupNonnativevsNative:Task2vs1	-0.10	0.02	-5.48	< .001
GroupNonnativevsNative:Task3vs1	-0.04	0.02	-2.14	.033

*Note: Model specification in lmer(CV ~ Group*Task +(1|SubjectNo)).*

Table A4. Estimates of performance of native and nonnative groups in the vocabulary size test

	β	***SE***	***t***	***p***
(Intercept)	167.61	2.82	59.33	< .001
GroupNonnativevsNative	-74.61	5.65	-13.21	< .001

Note: Model specification in lm(Vocab ~ Group).

Table A5. Fixed-effect estimates of accuracy performance of nonnative groups in the three processing tasks

	β	***SE***	***z***	***p***
(Intercept)	2.75	0.13	20.51	< .001
GroupSA-onsetvsAH-regular	-0.88	0.12	-7.10	< .001
GroupSA-onsetvsAH-intensive	-0.15	0.13	-1.19	.234
Task2vs1	-0.54	0.30	-1.80	.072
Task3vs1	0.10	0.32	0.30	.762
GroupSA-onsetvsAH-regular:Task2vs1	0.45	0.16	2.77	.006
GroupSA-onsetvsAH-intensive:Task2vs1	0.22	0.17	1.29	.198
GroupSA-onsetvsAH-regular:Task3vs1	-0.33	0.19	-1.70	.089
GroupSA-onsetvsAH-intensive:Task3vs1	0.22	0.21	1.03	.301

*Note: Model specification in is glmer(Accuracy ~ Group*Task + (1+Group|Item_number) + (1+Task|SubjectNo)).*

Table A6. Fixed-effect estimates of RT performance of nonnative groups in the three processing tasks

	β	***SE***	***t***	***p***
(Intercept)	3.14	0.47	6.72	< .001
GroupSA-onsetvsAH-regular	0.10	0.03	4.02	< .001
GroupSA-onsetvsAH-intensive	-0.00	0.03	-0.10	.922
Task2vs1	0.28	0.08	3.36	.001
Task3vs1	0.20	0.09	2.25	.026
log_audio_duration	0.65	0.07	9.80	< .001
Trial_number	-0.00	0.00	-7.58	< .001
GroupSA-onsetvsAH-regular:Task2vs1	-0.08	0.03	-2.76	.007
GroupSA-onsetvsAH-intensive:Task2vs1	-0.05	0.03	-1.55	.124
GroupSA-onsetvsAH-regular:Task3vs1	0.03	0.03	1.11	.267
GroupSA-onsetvsAH-intensive:Task3vs1	-0.04	0.03	-1.33	.186

*Note: Model specification in lmer(log_RT ~ Group*Task + log_audio_duration + Trial_number+ (1+Task|SubjectNo) + (1+Group|Item_number)).*

Table A7. Fixed-effect estimates of CV performance of nonnative groups in the three processing tasks

	β	***SE***	***t***	***p***
(Intercept)	0.33	0.00	90.41	< .001
GroupSA-onsetvsAH-regular	0.02	0.01	1.68	.095
GroupSA-onsetvsAH-intensive	0.01	0.01	1.36	.175
Task2vs1	0.04	0.01	6.15	< .001
Task3vs1	-0.07	0.01	-10.53	< .001
GroupSA-onsetvsAH-regular:Task2vs1	-0.03	0.02	-1.94	.053
GroupSA-onsetvsAH-intensive:Task2vs1	0.01	0.02	0.45	.653
GroupSA-onsetvsAH-regular:Task3vs1	0.00	0.02	0.22	.826
GroupSA-onsetvsAH-intensive:Task3vs1	0.02	0.02	1.00	.316

*Note: Model specification in lmer(CV ~ Group*Task +(1|SubjectNo)).*

Table A8. Estimates of performance of nonnative groups in the vocabulary size test

	β	***SE***	***t***	***p***
(Intercept)	131.47	3.26	40.34	< .001
GroupAH-regularvsSA-onset	-16.19	4.47	-3.62	< .001
GroupAH-intensivevsSA-onset	15.27	4.65	3.28	.001

Note: Model specification in lm(Vocab ~ Group).

Appendix B (Chapter 2)

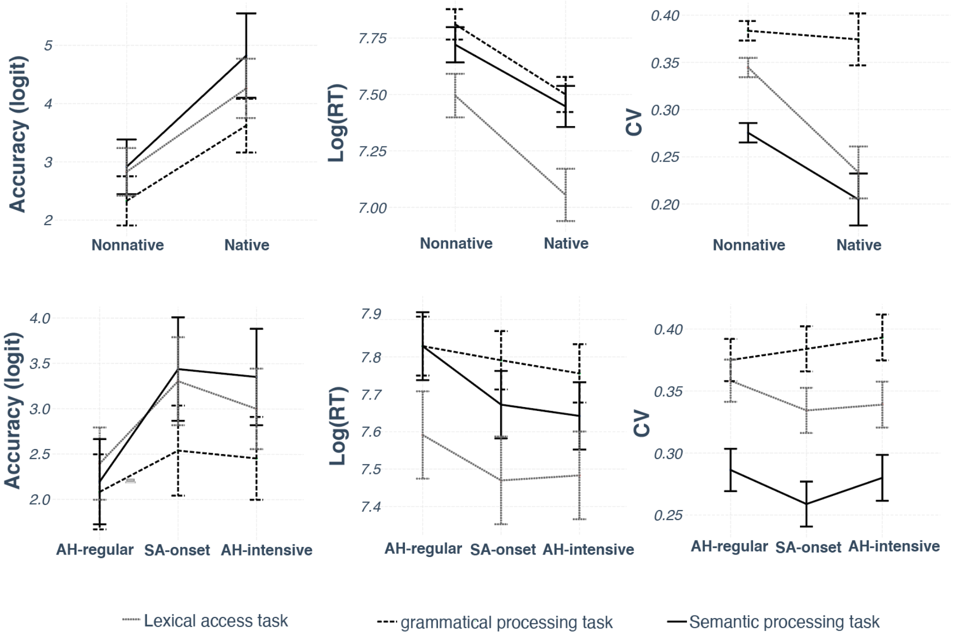

Figure B1. Interaction Between Group and Task Effects in the Processing Efficiency Models

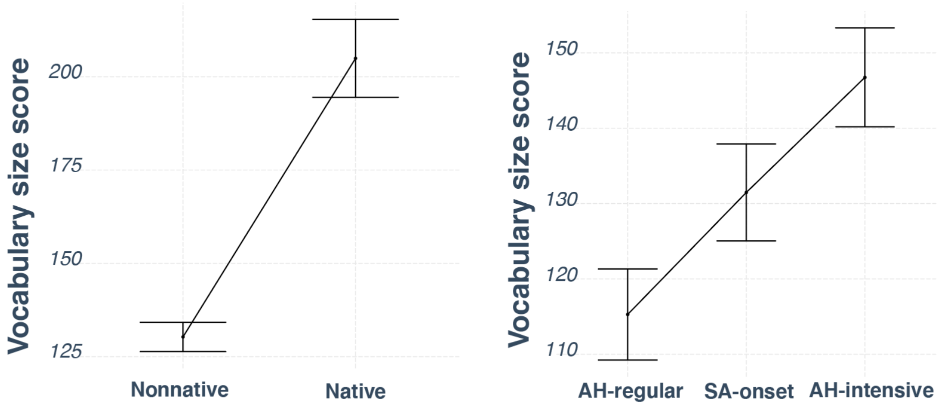

Figure B2. Group Effects in the Vocabulary Size Models

Appendix C (Chapter 2)

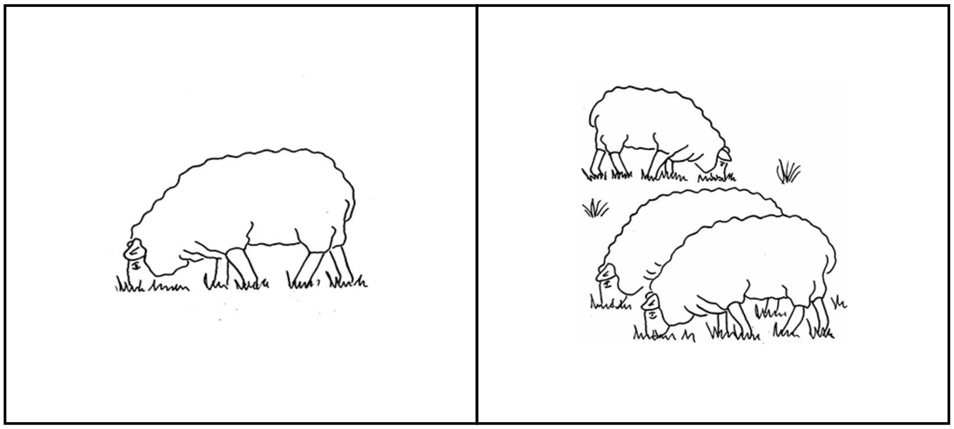

Figure C1. Sample visual display of the Grammatical Processing Task. Morphological-cue example: “the sheep eats”. (A) Correct picture. (B) Wrong picture.

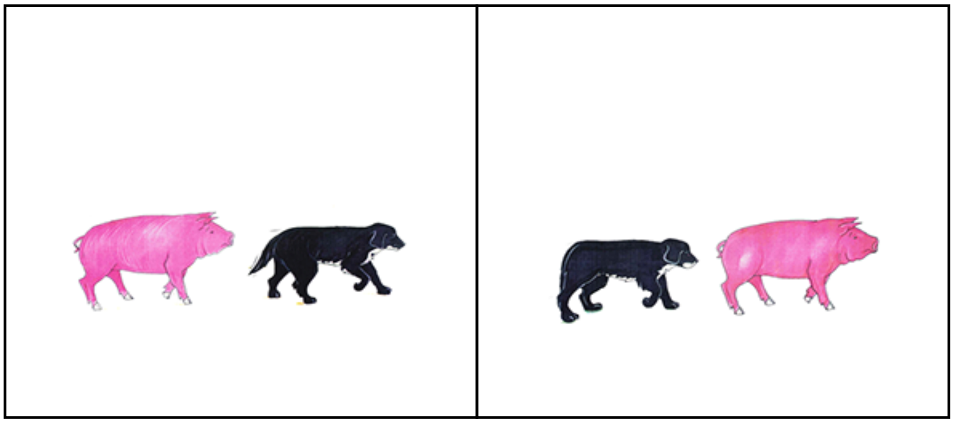

Figure C2. Sample visual display of the Grammatical Processing Task. Syntactical-cue example: “It is the dog that the pig follows”. (A) Correct picture. (B) Wrong picture.

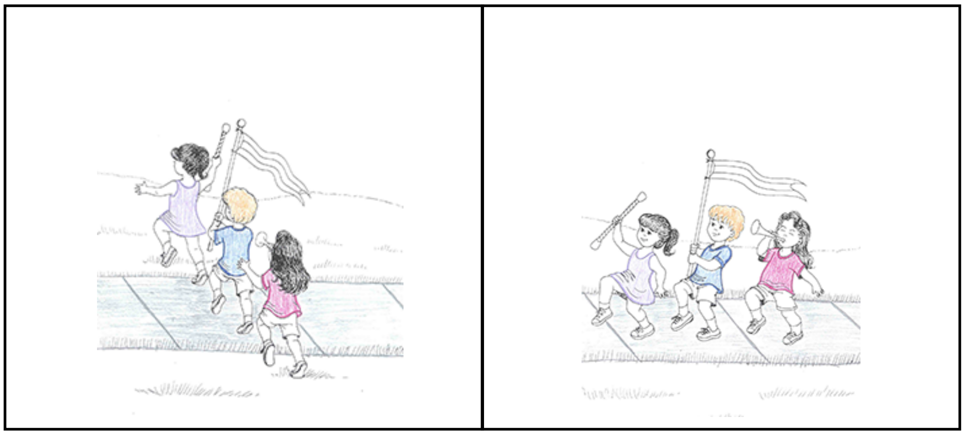

Figure C3. Sample visual display of the Grammatical Processing Task. Lexical-cue example: “The children are marching along the sidewalk”. (A) Wrong picture. (B) Correct picture.

Appendix D (Chapter 3)

Table D1. Fixed-effect estimates of vocab models (split by groups).

	β	*SE*	*t*	*p*
AH-regular group				
(Intercept)	116.45	3.77	30.91	< .001
TimePostvsPre	4.87	2.29	2.13	.038
AH-intensive group				
(Intercept)	147.54	2.45	60.31	< .001
TimePostvsPre	1.35	1.49	0.91	.369
SA group				
(Intercept)	131.57	3.27	40.24	< .001
TimePostvsPre	7.45	1.80	4.14	< .001

Note: Model specification in lmer(Vocab ~ Time + (1|SubjectNo)).

Table D2. Fixed-effect estimates of RT models (split by tasks).

	β	*SE*	*t*	*p*
Lexical access task				
(Intercept)	5.91	1.12	5.27	< .001
TimePostvsPre	-0.12	0.00	-29.97	< .001
GroupAH-regularvsSA	0.12	0.03	3.60	< .001
GroupAH-intensivevsSA	0.04	0.03	1.21	.228
log_audio_duration	0.18	0.18	1.00	.321
Trial_number	0.00	0.00	0.92	.359
TimePostvsPre:GroupAH-regularvsSA	-0.07	0.01	-7.02	< .001
TimePostvsPre:GroupAH-intensivevsSA	0.02	0.01	1.99	.047
Grammatical processing task				
(Intercept)	2.56	0.61	4.19	< .001
TimePostvsPre	-0.11	0.00	-27.25	< .001
GroupAH-regularvsSA	0.04	0.02	1.50	.136
GroupAH-intensivevsSA	-0.02	0.02	-0.69	.492
log_audio_duration	0.75	0.08	9.09	< .001
Trial_number	-0.00	0.00	-14.13	< .001
TimePostvsPre:GroupAH-regularvsSA	0.02	0.01	1.97	.049
TimePostvsPre:GroupAH-intensivevsSA	0.04	0.01	3.78	< .001
Semantic processing task				
(Intercept)	4.29	0.87	4.93	< .001
TimePostvsPre	-0.11	0.00	-31.00	< .001
GroupAH-regularvsSA	0.16	0.03	6.01	< .001
GroupAH-intensivevsSA	-0.01	0.03	-0.48	.629
log_audio_duration	0.50	0.12	4.33	< .001
Trial_number	-0.00	0.00	-7.08	< .001
TimePostvsPre:GroupAH-regularvsSA	-0.03	0.01	-3.04	.002
TimePostvsPre:GroupAH-intensivevsSA	0.03	0.01	3.08	.002

*Note: Model specification in lmer(log_RT ~ Time*Group + log_audio_duration + Trial_number+ (1|SubjectNo) + (1+Group|Item_number).*

Appendix E (Chapter 4)

Additional information on participants

Participants' scores on standardized tests, such as the Test for English Majors - Band 8 (TEM-8), College English Test – Band 6 (CET-6) and International English Language Testing System (IELTS) test, were also collected with the background questionnaire. Non-English-major students who scored lower than 500 in CET-6 turned out to have very low performance in the sentence comprehension tasks in our pilot study. As our focus was on relatively automatic language processing, which can only be investigated sensibly in learners with at least intermediate language proficiency, we decided to only include students with CET-6 scores above 500 (CET-6 scores form a normal distribution with 500 as the mean score). More than half of the people who responded to our advertisement were excluded for this reason. A large portion of the study-abroad learners who signed up had arrived in London two or three months earlier than other students to attend a pre-sessional language course to boost their slightly inadequate English proficiency. To minimize additional noise caused by different arrival times, we only invited newly-arrived students who had not entered the UK more than one month prior to taking our tests. Moreover, with regard to initial language proficiency before our pretest, we compared the CET-6 scores of our participant groups (involving 130 participants for whom CET-6 scores were available out of the total 143) with an ANOVA test. Results show that there were significant between-group differences in CET-6 scores (F(2, 127) = 7.22, p = .001). Post-hoc comparisons using the Tukey HSD test indicated that the SA group (M = 527.64, SD = 54.70) was significantly different from the AH-intensive group (M = 565.09, SD = 40.85) but not from the AH-regular group (M = 541.51, SD = 33.83). However, these results should be interpreted with caution because participants took the CET-6 test up to three years before the pre-test, and hence maybe well before the SA group started to prepare for studying abroad.

Appendix F (Chapter 4)

Table F1. Correlation matrix of observed individual-difference variables.

		1	2	3	4	5	6	7	8
1	LLAMA_B								
2	LLAMA_D	.18*							
3	LLAMA_F	.33*	.07						
4	Forward digit span	.22*	.23*	.14					
5	Backward digit span	.17*	.16	.13	.50				
6	Engagement	-.19*	-.13	-.15	.01	-.14			
7	Mental Well-being	-.02	-.01	-.04	-.02	-.03	.16		
8	Speaking time	-.01	.01	.08	.03	-.08	.44*	.08	

Note: * *p* < .05

Table F2. Model comparisons for evaluating whether models with higher-order interactions have better model fits than models without (higher-order) interactions. N-way models included all possible n-way interactions, lower-order interactions, and main effects of fixed-effect factors. Individual-difference variables were, however, not allowed to interact with each other to reduce the risk of overfitting models. For example, the fixed structure of the ACC 4-way model is the following: ACC ~ Group*Time*Task *(Aptitude + Working memory + Exposure + Social interaction + Emotion). This model is compared to a 3-way model which only contains 3-way interactions (as well as two-way interactions and main effects).

Model	***Df***	***AIC***	***BIC***	***logLik***	***deviance***	***ΔDf***	***Δchisq***	***p***
Vocab models								
Vocab_1way	11	2430.2	2470.4	-1204.1	2408.2			
Vocab_2way	28	2441.9	2544.1	-1192.9	2385.9	17	22.4	.171
Vocab_3way	38	2454.5	2593.2	-1189.3	2378.5	10	7.4	.691
ACC models								
ACC_1way	13	25884	25998	-12929	25858			
ACC_2way	46	25806	26207	-12857	25714	33	144.6	<.001 ***
ACC_3way	90	25839	26625	-12830	25659	44	54.8	0.127
ACC_4way	110	25864	26825	-12822	25644	20	14.9	0.783
RT models								
RT_1way	33	-723.50	-439.84	394.75	-789.50			
RT_2way	66	-712.31	-144.99	422.15	-844.31	33	54.8	.010

RT_3way	110	-763.65	181.87	491.83	-983.65	44	139.3	<.001 ***
RT_4way	130	-857.38	260.06	558.69	-1117.38	20	133.7	<.001 ***
CV models								
CV_1way	22	-2314.5	-2209.9	1179.2	-2358.5			
CV_2way	55	-2316.3	-2054.8	1213.1	-2426.3	33	67.8	<.001 ***
CV_3way	99	-2278.1	-1807.4	1238.0	-2476.1	44	49.8	.253
CV_4way	119	-2259.4	-1693.6	1248.7	-2497.4	20	21.3	.381

Table F3. Fixed-effect estimates of performance of participant groups in the vocabulary size test.

	β	***SE***	***t***	***p***
(Intercept)	133.78	1.78	75.24	< .001
TimePrevsPost	4.63	1.16	4.01	< .001
GroupAH-regularvsSA	-19.72	6.15	-3.20	.002
GroupAH-intensivevsSA	13.22	5.64	2.34	.021
Aptitude	9.84	2.75	3.57	< .001
Working memory	-2.62	2.07	-1.27	.207
Exposure	1.44	1.21	1.19	.235
Social Interaction	-0.48	1.09	-0.44	.658
Emotion	0.14	0.99	0.14	.887

Note: Model specification: lmer(Vocab ~ Time + Group + Aptitude + Working memory + Exposure +Social interaction+ Emotion +(1|SubjectNo))

Table F4. Fixed-effect estimates of accuracy performance of participant groups in the three language processing tasks.

	β	***SE***	***z***	***p***
(Intercept)	2.67	0.14	18.98	< .001
TimePrevsPost	0.20	0.04	4.56	< .001
GroupAH-regularvsSA	-0.62	0.20	-3.04	.002
GroupAH-intensivevsSA	0.21	0.18	1.16	.248

Task2vs1	-0.51	0.29	-1.73	.085
Task3vs1	0.15	0.31	0.48	.634
Aptitude	0.36	0.07	4.87	< .001
Working memory	-0.06	0.05	-1.19	.235
Exposure	0.07	0.03	1.98	.048
Social interaction	0.02	0.03	0.76	.447
Emotion	-0.00	0.03	-1.19	.850
TimePrevsPost:GroupAH-regularvsSA	-0.22	0.12	-1.91	.056
TimePrevsPost:GroupAH-intensivevsSA	-0.04	0.11	-0.35	.730
TimePrevsPost:Task2vs1	0.17	0.10	1.68	.094
TimePrevsPost:Task3vs1	0.10	0.11	0.88	.377
TimePrevsPost:Aptitude	0.04	0.05	0.69	.489
TimePrevsPost:Working memory	-0.01	0.04	-0.34	.736
TimePrevsPost:Exposure	-0.03	0.02	-1.44	.149
TimePrevsPost:Social interaction	0.00	0.02	0.23	.817
TimePrevsPost:Emotion	-0.01	0.02	-0.70	.484
GroupAH-regularvsSA:Task2vs1	0.27	0.17	1.54	.124
GroupAH-intensivevsSA:Task2vs1	0.02	0.16	0.10	.921
GroupAH-regularvsSA:Task3vs1	-0.43	0.23	-1.87	.061
GroupAH-intensivevsSA:Task3vs1	0.21	0.22	0.96	.337
GroupAH-regularvsSA:Aptitude	0.13	0.14	0.90	.366
GroupAH-intensivevsSA:Aptitude	0.04	0.15	0.27	.791
GroupAH-regularvsSA:Working memory	-0.09	0.11	-0.86	.389
GroupAH-intensivevsSA:Working memory	-0.13	0.11	-1.18	.237

GroupAH-regularvsSA:Exposure	-0.00	0.07	-0.05	.959
GroupAH-intensivevsSA:Exposure	0.01	0.06	0.11	.911
GroupAH-regularvsSA:Social interaction	-0.08	0.07	-1.10	.272
GroupAH-intensivevsSA:Social interaction	0.06	0.07	0.77	.443
GroupAH-regularvsSA:Emotion	0.08	0.05	1.47	.141
GroupAH-intensivevsSA:Emotion	0.10	0.05	1.89	.058
Task2vs1:Aptitude	0.03	0.08	0.34	.736
Task3vs1:Aptitude	0.22	0.10	2.15	.032
Task2vs1:Working memory	0.04	0.06	0.73	.469
Task3vs1:Working memory	0.01	0.08	0.15	.882
Task2vs1:Exposure	0.04	0.03	1.06	.288
Task3vs1:Exposure	0.05	0.04	1.12	.265
Task2vs1:Social interaction	-0.06	0.03	-2.07	.038
Task3vs1:Social interaction	-0.04	0.04	-1.05	.296
Task2vs1:Emotion	-0.05	0.03	-1.70	.090
Task3vs1:Emotion	-0.08	0.04	-2.26	.024

*Note: model specification: glmer(Accuracy ~ Time*Group + Time*Task+ Time*(Aptitude + Working memory + Exposure + Social interaction + Emotion) +Group*Task + Group*(Aptitude + Working memory + Exposure + Social interaction + Emotion)+ Task*(Aptitude + Working memory + Exposure + Social interaction + Emotion) +(1+Time+Task|SubjectNo) +(1+Time|Item_number))*

Table F5. Fixed-effect estimates of RT performance of participant groups in the three processing tasks (data split by task).

	β	*SE*	*t*	*p*
Lexical access task				
(Intercept)	4.88	0.95	5.16	< .001
TimePrevsPost	-0.14	0.03	-5.05	< .001
GroupAH-regularvsSA	0.21	0.08	2.72	.008
GroupAH-intensivevsSA	0.05	0.06	0.79	.432
Aptitude	-0.04	0.03	-1.68	.095
Working memory	0.02	0.02	1.12	.267
Exposure	0.00	0.01	0.30	.765
Social interaction	-0.00	0.01	-0.32	.752
Emotion	-0.00	0.01	-0.49	.627
log_audio_duration	0.34	0.15	2.29	.026
TimePrevsPost:GroupAH-regularvsSA	-0.18	0.07	-2.49	.014
TimePrevsPost:GroupAH-intensivevsSA	-0.04	0.06	-0.65	.519
TimePrevsPost:Aptitude	-0.02	0.02	-0.90	.370
TimePrevsPost:Working memory	0.00	0.02	0.16	.873
TimePrevsPost:Exposure	-0.02	0.01	-1.34	.183
TimePrevsPost:Social interaction	-0.01	0.01	-0.51	.609
TimePrevsPost:Emotion	0.01	0.01	0.69	.490
GroupAH-regularvsSA:Aptitude	-0.07	0.06	-1.13	.261
GroupAH-intensivevsSA:Aptitude	-0.03	0.06	-0.43	.665
GroupAH-regularvsSA:Working memory	0.01	0.04	0.24	.812
GroupAH-intensivevsSA:Working memory	0.02	0.05	0.50	.617
GroupAH-regularvsSA:Exposure	0.04	0.03	1.41	.162
GroupAH-intensivevsSA:Exposure	0.02	0.03	0.91	.363
GroupAH-regularvsSA:Social interaction	-0.02	0.03	-0.55	.580
GroupAH-intensivevsSA:Social interaction	-0.03	0.03	-0.94	.350
GroupAH-regularvsSA:Emotion	0.04	0.02	1.60	.112
GroupAH-intensivevsSA:Emotion	0.02	0.02	0.71	.479
TimePrevsPost:GroupAH-regularvsSA:Aptitude	0.08	0.06	1.39	.167

TimePrevsPost:GroupAH-intensivevsSA:Aptitude	-0.02	0.06	-0.29	.771
TimePrevsPost:GroupAH-regularvsSA:Working memory	-0.01	0.04	-0.21	.832
TimePrevsPost:GroupAH-intensivevsSA:Working memory	-0.01	0.04	0.17	.867
TimePrevsPost:GroupAH-regularvsSA:Exposure	-0.02	0.03	-0.53	.596
TimePrevsPost:GroupAH-intensivevsSA:Exposure	-0.01	0.03	-0.20	.845
TimePrevsPost:GroupAH-regularvsSA:Social interaction	0.02	0.03	0.58	.566
TimePrevsPost:GroupAH-intensivevsSA:Social interaction	0.02	0.03	0.68	.496
TimePrevsPost:GroupAH-regularvsSA:Emotion	-0.04	0.02	-1.81	.072
TimePrevsPost:GroupAH-intensivevsSA:Emotion	-0.01	0.02	-0.29	.772
Grammatical processing task				
(Intercept)	2.74	0.62	4.40	< .001
TimePrevsPost	-0.13	0.02	-6.41	< .001
GroupAH-regularvsSA	0.01	0.06	0.18	.856
GroupAH-intensivevsSA	-0.05	0.05	-1.10	.275
Aptitude	-0.04	0.02	-2.18	.031
Working memory	0.00	0.01	0.17	.868
Exposure	-0.02	0.01	-2.36	.020
Social interaction	0.00	0.01	0.24	.808
Emotion	0.00	0.01	0.30	.765
log_audio_duration	0.71	0.08	8.52	< .001
TimePrevsPost:GroupAH-regularvsSA	0.06	0.05	1.10	.275
TimePrevsPost:GroupAH-intensivevsSA	0.07	0.04	1.71	.090
TimePrevsPost:Aptitude	-0.03	0.02	-1.74	.085
TimePrevsPost:Working memory	0.02	0.01	1.56	.122
TimePrevsPost:Exposure	0.01	0.01	0.71	.479
TimePrevsPost:Social interaction	-0.01	0.01	-0.67	.505
TimePrevsPost:Emotion	-0.00	0.01	-0.21	.835

GroupAH-regularvsSA:Aptitude	-0.01	0.04	-0.32	.747
GroupAH-intensivevsSA:Aptitude	0.03	0.04	0.76	.452
GroupAH-regularvsSA:Working memory	0.02	0.03	0.52	.603
GroupAH-intensivevsSA:Working memory	-0.02	0.03	-0.76	.446
GroupAH-regularvsSA:Exposure	0.01	0.02	0.34	.738
GroupAH-intensivevsSA:Exposure	0.01	0.02	0.38	.704
GroupAH-regularvsSA:Social interaction	0.01	0.02	0.34	.731
GroupAH-intensivevsSA:Social interaction	-0.00	0.02	-0.20	.838
GroupAH-regularvsSA:Emotion	0.01	0.02	0.41	.681
GroupAH-intensivevsSA:Emotion	-0.01	0.02	-0.33	.743
TimePrevsPost:GroupAH-regularvsSA:Aptitude	0.03	0.04	0.72	.470
TimePrevsPost:GroupAH-intensivevsSA:Aptitude	-0.03	0.04	-0.61	.543
TimePrevsPost:GroupAH-regularvsSA:Working memory	-0.00	0.03	-0.13	.893
TimePrevsPost:GroupAH-intensivevsSA:Working memory	0.04	0.03	1.16	.247
TimePrevsPost:GroupAH-regularvsSA:Exposure	0.01	0.02	0.54	.591
TimePrevsPost:GroupAH-intensivevsSA:Exposure	-0.00	0.02	-0.23	.818
TimePrevsPost:GroupAH-regularvsSA:Social interaction	-0.04	0.02	-1.86	.066
TimePrevsPost:GroupAH-intensivevsSA:Social interaction	-0.12	0.02	-0.83	.408
TimePrevsPost:GroupAH-regularvsSA:Emotion	-0.00	0.02	-0.23	.817
TimePrevsPost:GroupAH-intensivevsSA:Emotion	0.00	0.01	0.02	.985
Semantic processing task				
(Intercept)	2.82	0.84	3.33	.002
TimePrevsPost	-0.14	0.02	-7.50	< .001
GroupAH-regularvsSA	0.06	0.06	1.04	.300
GroupAH-intensivevsSA	-0.08	0.05	-1.60	.111
Aptitude	-0.05	0.02	-2.52	.013

Working memory	0.01	0.01	0.39	.700
Exposure	-0.02	0.01	-2.37	.019
Social interaction	-0.00	0.01	-0.42	.674
Emotion	0.00	0.01	0.17	.868
log_audio_duration	0.69	0.11	6.19	< .001
TimePrevsPost:GroupAH-regularvsSA	-0.02	0.05	-0.39	.700
TimePrevsPost:GroupAH-intensivevsSA	0.05	0.04	1.30	.197
TimePrevsPost:Aptitude	0.00	0.02	0.09	.926
TimePrevsPost:Working memory	0.00	0.01	0.06	.949
TimePrevsPost:Exposure	0.00	0.01	0.27	.785
TimePrevsPost:Social interaction	-0.01	0.01	-0.76	.448
TimePrevsPost:Emotion	0.00	0.01	0.69	.491
GroupAH-regularvsSA:Aptitude	-0.04	0.04	-0.80	.426
GroupAH-intensivevsSA:Aptitude	-0.00	0.05	-0.07	.947
GroupAH-regularvsSA:Working memory	0.04	0.03	1.13	.261
GroupAH-intensivevsSA:Working memory	0.02	0.03	0.71	.480
GroupAH-regularvsSA:Exposure	-0.03	0.02	-1.12	.267
GroupAH-intensivevsSA:Exposure	0.00	0.02	0.05	.962
GroupAH-regularvsSA:Social interaction	0.03	0.02	1.53	.129
GroupAH-intensivevsSA:Social interaction	0.02	0.02	0.97	.334
GroupAH-regularvsSA:Emotion	0.01	0.02	0.63	.527
GroupAH-intensivevsSA:Emotion	0.00	0.02	0.03	.979
TimePrevsPost:GroupAH-regularvsSA:Aptitude	0.04	0.04	0.96	.341
TimePrevsPost:GroupAH-intensivevsSA:Aptitude	0.03	0.04	0.68	.499
TimePrevsPost:GroupAH-regularvsSA:Working memory	-0.01	0.03	-0.51	.614
TimePrevsPost:GroupAH-intensivevsSA:Working memory	-0.01	0.03	-0.35	.729
TimePrevsPost:GroupAH-regularvsSA:Exposure	-0.00	0.02	-0.08	.940
TimePrevsPost:GroupAH-intensivevsSA:Exposure	0.00	0.02	0.06	.954

TimePrevsPost:GroupAH-regularvsSA:Social interaction	-0.04	0.02	-1.87	.064
TimePrevsPost:GroupAH-intensivevsSA:Social interaction	-0.02	0.02	-0.99	.324
TimePrevsPost:GroupAH-regularvsSA:Emotion	0.01	0.01	0.69	.491
TimePrevsPost:GroupAH-intensivevsSA:Emotion	0.01	0.01	0.53	.599

*Notes: 1. Model specification: lmer(log_RT ~ Time*Group*(Aptitude + Working memory + Exposure +Social interaction + Emotion) + log_audio_duration +(1+ Time|SubjectNo) + (1 + Time|Item_number)*

2. Log_audio_duration was included in the model as a fixed-effect control variable to account for the potential effect of the audio duration of test stimuli on participants' response times.

Table F6. Fixed-effect estimates of CV performance of the participant groups in the three processing tasks.

	β	*SE*	*t*	*p*
(Intercept)	0.32	0.01	55.28	< .001
TimePrevsPost	-0.02	0.00	-5.31	< .001
GroupAH-regularvsSA	0.02	0.02	0.96	.341
GroupAH-intensivevsSA	0.03	0.01	2.24	.027
Task2vs1	0.06	0.01	9.77	< .001
Task3vs1	-0.07	0.01	-11.73	< .001
Aptitude	-0.00	0.01	-0.69	.495
Working memory	0.00	0.00	0.44	.662
Exposure	-0.00	0.00	-0.45	.657
Social interaction	0.00	0.00	0.53	.594
Emotion	0.00	0.00	0.49	.628
TimePrevsPost:GroupAH-regularvsSA	0.00	0.01	0.23	.821
TimePrevsPost:GroupAH-intensivevsSA	0.01	0.01	0.55	.587
TimePrevsPost:Task2vs1	0.04	0.01	4.91	< .001
TimePrevsPost:Task3vs1	0.01	0.01	1.07	.284

TimePrevsPost:Aptitude	-0.00	0.01	-0.32	.747
TimePrevsPost:Working memory	0.00	0.00	0.87	.388
TimePrevsPost:Exposure	0.00	0.00	1.15	.251
TimePrevsPost:Social interaction	-0.00	0.00	-0.98	.329
TimePrevsPost:Emotion	-0.00	0.00	-0.02	.983
GroupAH-regularvsSA:Task2vs1	-0.04	0.02	-1.91	.059
GroupAH-intensivevsSA:Task2vs1	-0.03	0.02	-1.41	.161
GroupAH-regularvsSA:Task3vs1	-0.01	0.02	-0.63	.530
GroupAH-intensivevsSA:Task3vs1	-0.01	0.02	-0.43	.665
GroupAH-regularvsSA:Aptitude	0.01	0.01	0.84	.405
GroupAH-intensivevsSA:Aptitude	0.02	0.01	1.62	.108
GroupAH-regularvsSA:Working memory	0.00	0.01	0.34	.737
GroupAH-intensivevsSA:Working memory	0.00	0.01	0.08	.940
GroupAH-regularvsSA:Exposure	-0.01	0.01	-0.90	.368
GroupAH-intensivevsSA:Exposure	-0.00	0.01	-0.48	.633
GroupAH-regularvsSA:Social interaction	-0.00	0.01	-0.70	.485
GroupAH-intensivevsSA:Social interaction	-0.00	0.01	-0.37	.710
GroupAH-regularvsSA:Emotion	-0.00	0.00	-0.91	.366
GroupAH-intensivevsSA:Emotion	0.00	0.00	0.52	.601
Task2vs1:Aptitude	0.01	0.01	1.12	.263
Task3vs1:Aptitude	-0.00	0.01	-0.10	.919
Task2vs1:Working memory	-0.00	0.01	-0.53	.600
Task3vs1:Working memory	0.00	0.01	0.58	.562
Task2vs1:Exposure	0.00	0.00	0.09	.931
Task3vs1:Exposure	-0.01	0.00	-1.62	.108
Task2vs1:Social interaction	-0.00	0.00	-0.31	.756
Task3vs1:Social interaction	0.00	0.00	0.07	.948
Task2vs1:Emotion	0.00	0.00	0.36	.718
Task3vs1:Emotion	-0.00	0.00	-0.26	.797

Note: Model specification: lmer(CV ~ Time*Group +Time*Task + Time*(Aptitude + Working memory + Exposure + Social interaction+ Emotion) + Group*Task +

Group*(Aptitude + Working memory + Exposure + Social interaction + Emotion) + Task*(Aptitude + Working memory + Exposure + Social interaction + Emotion) +(1+Time+Task|SubjectNo))

Appendix G (Chapter 5)

Table G1. Summary of included studies

Study_ID	Modality	Research focus	Groups (sample sizes)	L2 - L1	Participant Baseline proficiency	Duration of study	Instruments	Measures	Statistical analysis	Results (focusing on L2 gains over time)
Collentine (2004)	Speaking	Grammatical and lexical abilities	AH (20) vs. SA (26)	Spanish - English	- Intermediate (having at least two semesters of formal instruction)	16 weeks	Oral Proficiency Interview	- Broad grammatical measures (i.e., gender, number, person, mood, and tense). - 17 fine-grained grammatical accuracy measures (e.g., Copula, preposition); - narrative scores;	- ANCOVA for a broad analysis of gender, number, person, mood, and tense. - A discriminant analysis for grammatical variables.	- AH > SA in grammar - SA > AH in narrative scores

								- Information richness scores	- A discriminant analysis for lexical variables. -ANOVA for information richness	- AH > SA in lexical ability - SA > AH in information richness (i.e., semantic density)
Sasaki (2007)	Writing	Writing proficiency and fluency	AH (6) vs. SA (7)	English - Japanese	From low- to mid-intermediate	1 year (with 4-9 months of studying abroad)	- English proficiency test - Argumentative compositions	- English proficiency test scores - Composition scores - Writing fluency measures (i.e., quantity and speed)	ANOVAs	- SA = AH in general English proficiency - SA > AH in writing ability and fluency

Sasaki (2011)	Writing	Writing proficiency and fluency	AH (9) vs. SA-1.5-2 (9) vs. SA-4 (7) vs. SA-8-11 (12)	English - Japanese	First-year university student	3.5 years (1.5 to 11 months of studying abroad; with 4 testing points in total)	Argumentative compositions	Composition scores	ANOVAs	-SA > AH in writing abilities
Matsumura (2001)	Reading	Pragmatic competence	AH (102) vs. SA (97)	English - Japanese	University-level student	1 year (4 times of data collection)	Multiple-choice questionnaire (various scenarios)	Scores in scenarios for higher status, status equal, and lower status	SEM	SA > AH in pragmatic competence

Llanes & Muñoz (2013)	Speaking & Writing	Oral and written fluency, lexical and syntactic complexity, and accuracy	AH (20) vs. SA (46)	English - Spanish	University students	2 or 3 months	- Composition - Interview - Questionnaire	- Pruned syllables per minute / words per T-unit (depending on oral or written production) - Lexical richness - Clause per T-unit - Errors per T-unit	- Paired samples t tests for intragroup change over time - MANCOVA for intergroup differences on post-test performance	SA > AH in all oral variables AH > SA in written syntactic complexity
Hakanson & Norrby (2010)	Writing	Grammar and pragmatics	AH (9) vs. SA (11)	Swedish - English	Intermediate	8 months	- Translation, composition, communicative task, and interview for grammar - Gap-filling task for pragmatics	- Grammatical stages within Processability Theory - Scores of pragmatic competence	ANOVAs	- SA = AH in grammar - not explicitly reported for pragmatic results

							- Word association task for lexicon			

Serrano, Llanes, & Tragant (2011)	Writing & speaking	Written and oral fluency, syntactic and lexical complexity, and accuracy	-AH Intensive (69) vs. SA (24) for written productio n
			- AH Intensive (43) vs. SA (24) for oral productio n
			- AH Semi-Intensive (37) vs. SA (25) for written productio n
			- AH Semi-Intensive (12) vs. SA (25)
				English - Spanish	University students
					(AH intensive: Intermediate and advanced;
					AH semi-intensive: Intermediate;
					SA: no indicator of proficiency level)
						- 15 days for SA vs. AH Intensive.
						- 2 months for SA vs. AH Semi-Intensive.
							- Compositions
							- oral task (describing a picture)
								- Syllables per minute / words per T-unit (depending on oral or written production)
								- Clause per T-unit
								- Lexical richness
								- Errors per T-unit
									MANCOVAs	- SA = AH Intensive
										- SA > AH Semi-Intensive in fluency and lexical complexity

			for oral productio n							
Segalowitz & Freed (2004)	Speaking	Oral proficiency and fluency	AH (18) vs. SA (22)	Spanish - English	Having at least two semesters of formal instruction; No differences between groups at pre-test	1 semester	- Oral proficiency interview - lexical access task (semantic classification) - attention control	- Total words - Duration of speech - Number of words in longest run (turn) - Speech rate - Absence of hesitations/silent pause (hesit-free) - Absence of filled pause (filler-free) - Number of words in the longest fluent	- Two-way mixed ANOVAs for fluency measures - sign test and Chi-square analysis for OPI rating	SA > AH in turn, rate and filler-free

								speech run (fluent-run) - Rating of oral proficiency interview (ordinal measure)		
Freed, Segalowitz, & Dewey (2004)	Speaking	Oral proficiency and fluency	AH (18) vs. SA (22) vs. IM	French-English	-with 2-4 years of prior language instruction	- 12 weeks for SA - 7 weeks for IM - 12 weeks for AH	Oral Proficiency Interview	- Speech rate - Hesitation-free speech runs - Filler-free speech runs - Fluent runs - Repetition-free speech runs - Grammatical-repair-free speech runs - Total words spoken - Duration of speaking time	- ANOVAs	SA > AH in fluidity

								- Longest run		
Isabelli-Garcia (2010)	Reading	Grammar	AH (12) vs. SA (12)	Spanish-English	intermediate	4 months	- SOPI (ACTFL Proficiency Guidelines-Speaking) - Grammaticality judgment test	- Gender-marked attributive adjectival agreement - Non-gender-marked attributive adjectival agreement - Gender-marked predicative adjectival agreement - Non-gender-marked predicative adjectival agreement	- Kolmogorov-Smirnov statistical test - Wilcoxon Matched-Pairs Signed Rank test	SA = AH in gender agreement
Lafford (2004)	Speaking	Communication strategies (CS)	AH (20) vs. SA (26)	Spanish-English	intermediate	16 weeks	Oral Proficiency Interview	- Overall CS use	- Repeated-measure test	SA = AH in reduction of CS use(with repeated-measure test)

								- L1- and L2-based CS use - Direct and interactional CS use - Resource-deficit (RD), other-performance (OH), and own-performance (SF) CS use	- ANCOVA	SA > AH at post-test while controlling for pre-test (with ANCOVA)
Yu, Janse, & Schoonen (2020) (i.e., Chapter 3 of the present thesis)	Listening	Vocabulary size and processing efficiency	AH-regular (53) vs. SA (47) vs. AH-intensive (49)	English-Chinese	Intermediate-to-advanced	7 months	- PPVT vocabulary size test - lexical access task - grammatical processing task	- Vocabulary size - Processing accuracy - Processing speed	Mixed models	- SA = AH-regular > AH-intensive in vocabulary - AH-regular > SA > AH-intensive in processing efficiency

							- semantic processing task	- Processing stability		
Muñoz & Llanes (2014)	Speaking	Foreign accent	AH (12) vs. SA (15)	English-Spanish / Catalan	First-year university students	3 months	Picture-elicited narrative task	Foreign accent score	- ANOVAs - paired sample t test	SA > AH in foreign accent
Felix-Brasdefer & Hasler-Barker (2015)	Speaking	Pragmatics	AH (12) vs. SA (25)	Spanish - English	University students	8 weeks	Modified oral discourse completion task	Use of complement strategies	Paired-samples t test	- Both positive and negative pre-post changes in the SA group - No pre-post changes in the AH group
Montero, Serrano, & Llanes (2017)	Speaking	Communication strategies (CSs)	AH (24) vs. SA (22)	English – Spanish / Catalan	University students	3 months	Oral narrative task	- Effective CS - L1-based CS	Wilcoxon and Man-Whitney tests	- No pre-post changes for the SA group - A reduction of effective CS use for

										the AH group
Faretta-Stutenberg & Morgan-Short (2018)	Reading	Grammar	AH (29) vs. SA (20)	Spanish - English	University students	1 semester	Grammaticality judgment task	- GJT d' - GJT accuracy - EEG measures of processing change	Paired-samples t test	- both SA and AH groups made gains in accuracy - only the SA group achieved processing changes
Hirakawa, Shibuya, & Endo (2019)	Listening	Grammar	AH (13) vs. SA (12)	English - Japanese	Upper-elementary to low-intermediate	3 or 5 weeks	A preference task with audio and visual stimuli	Adjective ordering	Two-way repeated measures ANOVA	AH > SA
Cutrone & Datzman (2015)	Listening and reading	General proficiency	SA (44) vs. AH (27)	English-Japanese	Intermediate	3 or 3.5 weeks	TOEFL test	Listening, grammar, reading, and total scores	Paired samples t tests	AH > or = SA in general proficiency

Freed (1995)	Speaking	Oral proficiency and fluency	AH (10)* vs. SA (15)	French – English (mostly)	Varied	1 semester	Oral Proficiency Interview	Globe fluency and OPI scores	not compared with statistical tests	NA
Jochum (2014)	Speaking	Oral proficiency and fluency	AH (9) vs. SA (9)	English-Spanish	Intermediate	1 semester	Oral Proficiency Interview	OPI scores	ANOVA	SA > AH in oral proficiency

Supplementary materials (Chapter 5)

Data File 1

Filename: Appendix Table G2. Overview of effect sizes and moderators.xlsx

Description: This accompanying Excel spreadsheet shows effect sizes and variances of the included studies as well as how moderator variables were coded. In the column headings, “g.SA” and “g.var.SA” represent effect sizes in Hedge’s g and standard error for the pretest-to-posttest change for study-abroad groups; “g.AH” and “g.var.AH” for at-home groups; “g” and “g.var” for the difference in the effect between study-abroad and at-home Groups.

Data File 2

Filename: Appendix Table G3. Data extraction.xlsx

Description: This Excel spreadsheet shows the extracted parameters from the included studies for the calculation of effect sizes. These parameters are sample sizes (“ni”), means and standard deviations of both pretest and posttest performance (“m_pre”, “m_post”, “sd_pre”, “sd_post”), as well as the pretest-posttest correlations (“ri”) for each group on all measures.

Acknowledgments

This PhD project has been the most difficult yet rewarding experience in my life so far. I am very grateful to all the people that have helped me along the way to make the impossible possible.

First and foremost, I would like to thank my supervisors, Rob and Esther. Rob, I felt so lucky when you gave me the opportunity to join you and your research group in the beautiful Netherlands. Your amiable manner and witty remarks never failed to put me at ease. I greatly appreciate your generous support, which has benefited all aspects of my academic life and development. And thank you for always being patient with me when I got stuck, for giving me maximal freedom, and for providing insightful feedback on my writing. Esther, you were involved in this research project relatively late but very quickly proved to be an indispensable part of this team. I admire your sharp mind and eye for detail. Thank you for constantly challenging me to clarify my ideas and improve the flow of my argumentations. I miss our intense but always cheerful supervision meetings with the three of us, especially those face-to-face ones before the outbreak of COVID-19.

My sincere gratitude also goes to my manuscript committee members, i.e., Prof. Peter-Arno Coppen, Prof. Jean-Marc Dewaele, Prof. Jonas Granfelt, Prof. Marjolijn Verspoor, Dr. Sible Andringa, Dr. Elke Peters, and Dr. Marianne Starren. Thank you for evaluating my thesis and providing valuable feedback.

Lotte and Chen, thank you for being my paranymphs. Lotte, you are the best office mate one could have wished for, full of fun and always ready to help. A special thank you for helping me translate the summary of this book into Dutch. I enjoyed our daily casual chats, all your traveling tips that inspired me to make my own intra-European trips, as well as the abundant supply of sweets in your drawer. And I appreciate our regular catching-ups after I have left the Netherlands very much. I hope we will see each other again in the near future, and look forward to being your local guide in Nanjing! Chen, when you started your PhD, we were naturally drawn together. Despite the seemingly huge difference in our personalities, it has always been a great delight to hang out with you and it turned out we also shared many hobbies. Our friendship only grows as time passes by. You are a shining star, talented, energetic, and curious. I cannot wait to see the wonderful things the future has in store for you! Hope our earlier conversations and ideas can be developed into fruitful collaborations.

My data collection in the UK and China was a highlight of this project. I would like to express heartfelt thanks to my local hosts and short-term supervisors, Prof. Jean-Mac Dewaele and Prof. Dongmei Ma. Jean-Marc, thank you for warmly welcoming me into your research group, providing tremendous help with data collection, and involving me in very interesting academic as well as non-academic activities. I had a super great time in London! Ma (Laoshi), I cannot thank you enough for introducing the world of second language research to me during my Master's and motivating me into pursuing a

doctoral degree abroad afterward. It is due to your kindest help and support that I could run the experiments of this PhD project so successfully and smoothly in China. I was also grateful to all my participants who dedicated hours to participating in this research even when their own schedules were quite packed. Xinyue Wang and Yiting Di, thank you very much for administering the tests together with me. Without you, it would have taken me much more time and effort to finish collecting data. I am also indebted to many other people who helped me with practicalities (e.g., sending out recruitment advertisements and arranging for accommodations): Chengchen, Daiqun, Guoren, Heyi, Jinyan, Ming, Siyu, Yangmin, Yanqiu, Yiting Wu, Yuqi, Zhenzhu, Zhuwei, Ziyun, and so forth.

One of the privileges of doing a PhD at Radboud University Nijmegen is that we get to learn from top-notch researchers and have academic training provided by multiple excellent research institutes, i.e., Centre for Language Studies, Max Planck Institute for Psycholinguistics, and Donders Institute for Brain, Cognition, and Behaviour. I would like to thank my colleagues, especially (former) members of the Cognitive and Developmental Aspects of Multilingualism group. Chantal, Claire, Elly, Eva, Ferdy, Figen, Gert-Jan, Laurel, Limor, Saskia, Sharon, Stefan, Sybrine, Ton, and many others, thank you all for the inspiring presentations and discussion, which gradually shaped up my understanding of psycholinguistics. I want to give a big shout-out to Kevin for organizing many wonderful courses, workshops, and events, and for striving to cater to the professional development of every IMPRS student. I also want to send a big thank you to Louis. I appreciated very much your lmer course and your allowing me to pester you about statistics whenever I ran into problems! Margret and Bob, thank you for offering assistant with arranging for testing equipment, configuring testing computers, and troubleshooting. Peter, thank you for organizing useful workshops and events and for making paperwork with GSH so efficiently. I would also like to extend my gratitude to everyone on the 8th (and possibly also 9th floor) for making the working environment so friendly and fun: Aurora, Baiyu, Emily, Eric, Erwin, Hanno, Helmer, Henk, Katharine, Mario, Martijn, Micha, Nelleke, Tim, Theresa, Thijs, Wei, Wessel, Xing, Zhongnan and many others.

It is a very fortunate coincidence for me that my PhD project and SAREP network started around the same time. Being a member of this network, which is particularly pertinent to my research topic, has many benefits. Martin, thank you for bringing people in the subfield of study abroad research together and organizing amazing networking events! I also want to thank all the members of the SAREP network for interesting presentations, nice encounters, and the exchange of ideas. Special thanks go to Jonas, Marianne, Carmen, and Rosamond for organizing conferences, workshops, and book exhibitions.

Thank you to my friends in China, the Netherlands, and the UK. Huang, Jieying, Peng, and Xue, thank you for frequently hanging out with me and creating loads of happy memory! You enriched my social life when I was in the Netherlands, and we still listen to and support each other despite the long distance between us. Hongling, Keyang, Qi, Tao, Wei, Weibin, Yidong, Yueyue, and Yuxi, thank you for your kindest help and invitations to delicious homemade dinners when I first arrived in Nijmegen. GiGi, Junlei, Peter, Qihua, Shuang, Xuefei, and Xueli, thank you for letting me join your weekly fellowship events and outings. I appreciated the warm and relaxing atmosphere as well as the delicious and free Chinese dishes. Ziye, without you, I would have been homeless when I set foot in Nijmegen. You helped me establish my new life in a foreign country in a comfortable and effortless manner. Polina, you were my idol when I started my PhD, which you probably did not know. Thank you for letting me benefit from your knowledge of linguistics, statistics, the university, the Netherlands, and other stuff. I missed your iconic "how are you"s and wished that we had more time hanging out after work. Xue, I still remembered that afternoon when we met in a corridor. After that, a long chain of cheerful late-night eating, drinking, and chatting followed. If only we had known each other much earlier! And thank you for giving me a hand when I was about to leave the Netherlands. I am so excited about our future gatherings in Nanjing! Leilei, we have been best friends since high school and you have never let me down. I miss our European trip and hope we can travel to many more places in the world together. A special thank you for designing the cover of this book and some of the stimulus pictures for me. Zhaorong, thank you for often helping me and motivating me into becoming better. Abir, David, Hong, Jackie, Jie, Jiyuan, Jinbiao, Jing, Maozong, Nan, Shanshan, Ting, Xuanzhi, Xuzhao, Ying, Yongqiang, Yuan, Yuanyuan, and Zhaohui, I enjoyed the good time we spent together and hope our future will cross again.

Finally, I would like to take this opportunity to express my love to my parents and sisters for always making me feel loved. Thank you for taking pride in all my achievements, supporting me in all the ways you can, and believing in me no matter what.

Curriculum Vitae

Xiaoru Yu was born in 1989 in Henan, China. She obtained her Bachelor of Arts degree in English Education at Henan Normal University, China, in 2012. She then completed a Master of Arts degree in Foreign Linguistics and Applied Linguistics at Southeast University, China, in 2016. In the same year, she was awarded a scholarship by Chinese Scholarship Council to pursue a doctoral degree in Linguistics at Radboud University in the Netherlands. She carried out her research there between 2016 and 2021, as a part of both the Graduate School of Humanities and the International Max Planck Research School for Language Sciences.

List of publications

Publications:

Yu, X., Janse, E., & Schoonen, R. (2020). The effect of learning context on L2 listening development: Knowledge and processing. *Studies in Second Language Acquisition, 43*(2), 329–354.

Ma, D., Yu, X., & Zhang, H. (2017). Word-level and sentence-level automaticity in English as a foreign language (EFL) learners: A comparative study. *Journal of Psycholinguistic Research, 46*(6), 1471–1483.

Submitted manuscripts:

Yu, X., Janse, E., & Schoonen, R. (submitted). The impact of studying abroad on L2 development: A multilevel meta-analysis.

Yu, X., Janse, E., & Schoonen, R. (submitted). Individual differences and L2 listening development in study-abroad and at-home contexts.

Yu, X., Janse, E., & Schoonen, R. (submitted). Breaking down listening comprehension: How does L2 exposure affect vocabulary knowledge and processing efficiency in EFL learners.