ABSTRACT

Title of Document: USING NEW MEASURES OF IMPLICIT L2 KNOWLEDGE TO STUDY THE INTERFACE OF EXPLICIT AND IMPLICIT KNOWLEDGE

Yuichi Suzuki, Doctor of Philosophy, 2015

Directed By: Dr. Robert M. DeKeyser, Second Language Acquisition

Second language acquisition (SLA) becomes extremely difficult for late second language (L2) learners, who are assumed to have passed the sensitive or critical period for L2 learning. As one of the major accounts of the post-critical period L2 learning processes, a fundamental distinction between explicit and implicit learning and knowledge was proposed over three decades ago. The first goal of the current study was to develop fine-grained measures for implicit knowledge to distinguish it from automatized explicit knowledge. The second goal was to use these validated measures to explore the interface issue of explicit and implicit knowledge by correlating these measures with several cognitive aptitudes.

One hundred advanced L2 Japanese speakers whose first language was Chinese were recruited; they were given tests for both automatized explicit knowledge and implicit knowledge, along with three cognitive aptitude measures. The present study developed three psycholinguistic tasks that can reliably assess implicit knowledge (the eye-tracking-while-listening task, the word-monitoring task, and the self-paced reading task) and compared them with the existing tasks that have been claimed to measure implicit knowledge (time-pressured form-focused tasks like grammaticality judgment tasks), but which we hypothesized tap into automatized explicit knowledge. The aptitude test battery consisted of LLAMA F, a measure of explicit learning

aptitude, the Serial-Reaction Time (SRT) task, a measure of implicit learning aptitude, and the letter-span task, a measure of phonological short-term memory.

In order to validate the measures for automatized explicit knowledge and implicit knowledge, a series of confirmatory factor analyses (CFA), multi-trait multi-method (MTMM) analyses, and structural equation model (SEM) analyses were conducted. Results confirmed that the existing tasks purported to measure implicit knowledge in fact tap into automatized explicit knowledge, whereas the new psycholinguistic measures tap into implicit knowledge. For the participants as a whole, the convergent validity for implicit knowledge measures was less than ideal. When the results were analyzed separately by length of residence, however, acceptable convergent validity for implicit knowledge was obtained for those with longer length of residence but not for those with shorter length of residence.

In order to address the interface issue, SEM analyses were conducted to investigate the relationship between automatized explicit knowledge and implicit knowledge. Results showed that automatized explicit knowledge significantly predicted the acquisition of implicit knowledge. Furthermore, the aptitude for explicit learning was the only significant predictor of the acquisition of automatized explicit knowledge, not for the acquisition of implicit knowledge. The effects of implicit learning aptitude and phonological short-term memory on the acquisition of both types of linguistic knowledge were limited.

In conclusion, the study demonstrated that the newer measures for implicit knowledge are more sensitive and opens up promising directions for developing additional fine-grained measures for implicit knowledge. The current findings provide the first empirical evidence at the latent construct level that automatized explicit knowledge, which develops through explicit learning mechanisms, impacts the acquisition of implicit knowledge.

USING NEW MEASURES OF IMPLICIT L2 KNOWLEDGE TO STUDY THE INTERFACE OF EXPLICIT AND IMPLICIT KNOWLEDGE

By

Yuichi Suzuki

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Robert M. DeKeyser, Chair
Professor Yi Ting Huang
Professor Steven J. Ross
Professor Nan Jiang
Professor Jeff MacSwan

Acknowledgements

After completing my master's program at the Tokyo Gakugei University, I wanted to investigate how L2 automatization takes place and how automaticity can be achieved more efficiently in L2 learning. Since then, my research interests have always come back to the issues revolving around automaticity. My PhD studies deepened my understanding of the things I wanted to know and broadened my view on L2 learning processes. I made the best decision to study SLA at UMD, and I am truly grateful to all the people I have met. They influenced me on many levels and helped me go through the journey with joy.

First and foremost, I would like to express my greatest appreciation to Dr. Robert DeKeyser, who supported me throughout the course of my study. I have always appreciated his responsive feedback, advice, and encouragements on my work. His insightful views on L2 learning fascinated me; for instance, his extensive understanding and foresight on the issue of automaticity, explicit and implicit learning/knowledge always shed light on the trail I decided to take. In the future, I would like to gain as broad expertise as he has on both theoretical and applied issues in SLA.

I would also like to thank my committee members, Dr. Yi Ting Huang, Dr. Steve Ross, Dr. Nan Jiang, and Dr. Jeff MacSwan for their support for my dissertation work. Dr. Yi Ting Huang became my main co-advisor with Robert; I would like to show my greatest appreciation to her for patiently training me in every detailed step of conducting an eye-tracking study. She taught me the nuts and bolts of psycholinguistic research, which is a treasure for me. Dr. Steve Ross always provided me insightful suggestions on statistical analyses for all of the projects I carried out at UMD. In the meetings, I often learned new statistical analyses and ingenious ways of carrying them out. I am also very grateful to Dr. Nan Jiang, who first introduced me to using

the word-monitoring task in his class project, which opened up my psycholinguistic line of research. My dissertation project is built on the expertise that they all brought to the study.

I also received generous support and guidance from Drs. Mike Long and Kira Gor, particularly when I was in my first year. The studies I conducted under their guidance created the foundation of my research skills, and I always appreciate their warm encouragements.

Special thanks go to Dr. Colin Phillips, who pushed me to go beyond my "comfort zone" and always gave me great advice at the right moments. IGERT indeed changed me and pushed me to go as far as I could. Without his advice and support, this work would not have been possible.

I would also like to thank to Drs. Cathy Doughty and Jared Linck for allowing me to use one of the measures from the Hi-LAB. Special thanks goes to Cathy for being on the committee for my first qualifying paper, and for her encouragements since then, and to Dr. Scott Barry Kaufman for allowing me to use his wonderful serial-reaction time task. Special thanks go to Sharon vonBergener and Maggie Kinser for painstakingly proofreading the draft of this dissertation. I also extend my gratitude to Kei Harata and Jun Fujita for their generous assistance in creating the stimuli sentences for this project.

Many people helped me conduct the intense data collection in Tokyo. My deepest appreciation goes to Dr. Yuki Hirose, who was so caring and provided all the support I needed during my stay at the University of Tokyo for my data collection. Without her support, this dissertation would not have been possible. My thanks also extend to all the members of her lab, who welcomed me and made my stay very enjoyable. I want to express my deepest gratitude to Drs. Kaoru Koyanagi, Yukiko Okuno, Tomomi Nishikawa, and Hiromi Ozeki, whose work on Japanese SLA has always inspired me and helped me recruit many L2 Japanese speakers near

Tokyo. I also want to thank Dr. Kiyoko Tadokoro and many Japanese teachers on the JASLA list serve, who spread the word of my current project to their students.

I would like to express my gratitude to the Integrative Graduate Education and Research Traineeship (IGERT) program of the National Science Foundation (NSF), the Office of the Graduate Dean for a Summer Research Fellowship, the PhD program in Second Language Acquisition at the University of Maryland, and the Language Learning Dissertation Grant Program for their financial support. Without their generous support, the extensive data collection would not have been possible.

I would like to thank my teachers at the Tokyo Gakugei University, who always welcome me whenever I go back to Tokyo. Prof. Ken Kanatani inspired me to study abroad and taught me how to carry out empirical research on English education. I also am indebted to Profs. Megumi Shimada, Hiroko Yabe, Yoshiki Takayama, Misato Usukura, Testuo Baba, and Etsuko Ota for their continuing support.

My special thanks extend to my friends and colleagues at the University of Maryland for their moral support and meaningful discussion on SLA. I would also like to express my sincerest gratitude to my family for their warm encouragements. Last but not least, this dissertation is dedicated to Ayami Suzuki, who always supported, encouraged, and believed in me.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Achieving high-level skills in a second language (L2) is a difficult task, and considerable variability is observed in Second Language Acquisition (SLA). SLA researchers have been trying to explain why first language acquisition is always successful, while for adult or late learners assumed to have passed the sensitive period(s) or optimal time window for L2 learning, SLA is not successful (Abrahamsson & Hyltenstam, 2008, 2009; DeKeyser, 2000; DeKeyser, Alfi-Shabtay, & Ravid, 2010; Flege, Yeni-Komshian, & Liu, 1999; Granena & Long, 2013; Johnson & Newport, 1989; Long, 2007). This central issue has occupied many researchers; many explanations have been proposed, ranging from biological differences to social and psychological conditions (see e.g., Hyltenstam & Abrahamsson, 2003). Among those, the distinction between explicit and implicit learning is an important element of explanation regarding variability in SLA.

The issue of implicit and explicit learning mechanisms has attracted attention from many SLA researchers because of its theoretical and educational implications (e.g., R. Ellis et al., 2009; Hulstijn, 2005). To tackle the issues surrounding explicit and implicit knowledge and learning, the methodological problem of measuring implicit knowledge is crucial. Previous SLA studies have shown empirically that explicit and implicit knowledge are distinct constructs that can be measured with different tests (Bowles, 2011; R. Ellis, 2005; Gutiérrez, 2013; Zhang, 2014). A recent study, however, employed word monitoring, a more fine-grained psycholinguistic technique to examine real-time predictive sentence processing and cast doubt on the validity of the existing implicit knowledge measures (Suzuki & DeKeyser, in press). The present dissertation attempts to develop a battery of fine-grained tests for implicit knowledge by applying various psycholinguistic methods to SLA.

With those more valid measures for implicit knowledge, the present study further aims to explore two related issues concerning the relationship between explicit and implicit knowledge (DeKeyser, 2007b; N. C. Ellis, 2005; R. Ellis, 2008; Hulstijn, 2002; Krashen, 1985; Paradis, 2009). First, it addresses the so-called interface issue, namely, does one type of knowledge lead to the other or does each necessarily develop independently? Little focused empirical research has been conducted to clarify the relationship between the two types of linguistic knowledge, largely due to the dearth of methods used to assess the two types of linguistic knowledge. The current study attempts to validate measures for automatized explicit knowledge and implicit knowledge, and it aims to investigate whether automatized explicit knowledge leads to the acquisition of implicit knowledge.

Second, the current study addresses how the mechanisms of explicit and implicit *learning processes* are deployed to acquire explicit and implicit knowledge. Most SLA researchers agree on the fact that both explicit and implicit learning are necessary for acquiring high levels of skill in L2 (N. C. Ellis, 2005; R. Ellis, 2004; Long, 2015; Paradis, 2009). The extent to which explicit and implicit learning contribute to adult L2 learning is yet to be determined. To address this question, the current study examines the relationship between different types of *linguistic knowledge* and *cognitive aptitudes*. Investigating interactions between explicit/implicit knowledge and cognitive aptitudes allows for shedding light on the nature of learning processes by making an inference about a mental process that is facilitated or hampered by individual differences in different aptitude components (DeKeyser, 2003, 2012; DeKeyser & Koeth, 2010). Investigating the role of aptitudes further adds pieces to the large puzzle of the interface issues of explicit and implicit knowledge.

Chapter 2: Review of the Literature

In this section, operationalization and measurements of explicit and implicit knowledge in previous studies will be evaluated critically, and limitations of the existing measures for implicit knowledge will be pointed out. To provide background for the development of a more fine-grained measure for implicit knowledge, an emerging line of investigations on online sentence processing in L1 and L2 psycholinguistics will be reviewed. The review of the psycholinguistically oriented SLA research opens up a promising avenue for tackling a core issue of SLA, i.e., the interface of implicit and explicit knowledge, by deploying multiple online measurements. The question of how explicit and implicit knowledge and learning related to each other will be reviewed critically. Background will also be provided on previous research on the role of cognitive aptitudes in ultimate attainment in SLA, to set the stage for our approach to examining the two types of learning mechanisms.

2.1 Validation of Explicit and Implicit Knowledge Measures in SLA

Despite the importance of the concepts of explicit and implicit knowledge and learning in the SLA field, only a few empirical studies have attempted to validate multiple measures of explicit and implicit knowledge. A seminal study conducted by the R. Ellis (2005) will be critically evaluated in this section. R. Ellis (2005) proposed seven criteria for the development of explicit and implicit knowledge tests (Table 1).

Table 1. *Operationalization of Explicit and Implicit Knowledge*

Criterion	Implicit knowledge	Explicit knowledge
Degree of awareness	Response according to feel	Response using rules
Time available	Time pressure	No time pressure
Focus of attention	Primary focus on meaning	Primary focus on form
Systematicity	Consistent responses	Variable responses
Certainty	High degree of certainty	Low degree of certainty
Metal. Knowledge	Metal. knowledge not required	Metal. knowledge required

Learnability	Early learning favored	Late, form-focused instruction favored

Four criteria are related to task design: degree of awareness; time available; focus of attention; and use of metalinguistic knowledge. The degree of awareness can be the primary, possibly sole, criterion for implicit knowledge (DeKeyser, 2009; Suzuki & DeKeyser, in press). This criterion refers to the extent to which learners are aware of their linguistic knowledge; the implicit knowledge measure has to assess whether or not L2 learners can use linguistic knowledge *without awareness*. The rest of the criteria are considered to be necessary conditions for tapping into implicit knowledge, because it is assumed that explicit and metalinguistic knowledge are less likely to be accessed when the task is carried out under time pressure and the test-taker's attention is focused on meaning. In other words, implicit knowledge is more likely to be drawn on under those conditions, but those are not sufficient conditions. When the stringent awareness criterion is applied, implicit knowledge can be distinguished from automatized explicit knowledge (Suzuki & DeKeyser, in press). Automatized explicit knowledge, even if its use if fast, involves use of linguistic knowledge *with* awareness; this point has been overlooked in previous research on explicit and implicit knowledge.

With the four task design criteria in mind, R. Ellis (2005) developed a test battery for explicit and implicit knowledge measures, consisting of an oral narrative task, a timed Grammaticality Judgment Test (GJT), an Elicited Imitation (EI) task, an untimed GJT, and a metalinguistic knowledge test (Table 2). The oral narrative task and the EI were predicted to measure primarily implicit knowledge because participants would be focused on meaning under time pressure. In contrast, the metalinguistic knowledge test was predicted to measure explicit knowledge because it asked participants to choose the best explanation of grammatical errors, which requires a high degree of awareness of and the use of meta-language. Both of the GJTs

required participants to focus attention on form, but the prediction was made that the timed GJT would draw on implicit knowledge more strongly, whereas the untimed GJT would tap into explicit knowledge because test-takers were given enough time to reflect and to use their metalinguistic knowledge to answer in the untimed test, but not in the timed test.

Table 2. *Design features of the tests*

Criterion Imitation	Oral Narrative	Elicited Imitation	Timed GJT	Untimed GJT	Metalinguistic
Degree of awareness	Feel	Feel	Feel	Rule	Rule
Time available	Pressured	Pressured	Pressured	Unpressured	Unpressured
Focus of attention	Meaning	Meaning	Form	Form	Form
Metalinguistic knowledge	No	No	No	Yes	Yes

The test battery was administered to 91 English L2 speakers (beginner to advanced proficiency), and the data was submitted to a confirmatory factor analysis (R. Ellis & Loewen, 2007).[1] The results supported the prediction: the three implicit knowledge measures (i.e., oral narrative, timed GJT, and EI) and the two explicit knowledge measures loaded onto separate factors. This finding was replicated in Bowles (2011), who tested both intermediate-level L2 Spanish learners (n=10) and Spanish heritage learners (n =10).[2] A more recent study also replicated the findings with a sample from Chinese first-year university students (Zhang, 2014).[3]

These studies empirically demonstrated that explicit and implicit knowledge could be measured separately to some extent, but their operationalization in the form of specific measurements is not without problems in four different respects: (1) assessment of awareness,

[1] A principal component factor analysis was initially conducted in R. Ellis (2005), but Isemonger (2007) criticized its use, on the grounds that the study afforded a prediction based on a model. Therefore, the confirmatory factor analysis was conducted in response to this criticism in Ellis and Loewen (2007).

[2] The number of participants in Bowles (2011) is apparently small; the results must be interpreted with caution.

[3] The CFA results showed that the correlation between the two factors was very high (r = .86), indicating the lack of discriminant validity.

(2) the criterion of certainty and systematicity, (3) a criterion of learnability, and (4) time pressure as a result of task design.

First, R. Ellis (2005) operationalized the most important criterion, the degree of awareness, by the retrospective report about whether they made use of *feelings* or *rules* in responding to a task. In the untimed GJT, participants were asked to indicate whether they judged the sentence based on feelings or rules for each test item. He then computed correlation coefficients between the learners' applications of rule in the untimed GJT (a percentage score based on the participants' reported use of rule) and the other test scores. It was hypothesized that only the untimed GJT and the metalinguistic knowledge test would correlate with the use of rules. The results showed that rule use was significantly related to the ungrammatical items on the untimed GJT and the metalinguistic knowledge test. This approach assumes distant links from the use of rule in one task to the other test scores, which does not offer direct evidence for the use of rule in the other tasks. Moreover, no correlations were reported for the use of feel with other test scores. Overall, relying on self-report of awareness makes it very difficult to draw firm conclusions about their awareness in the use of linguistic knowledge.

In addition, three other criteria in Table 1 (systematicity, certainty, and learnability) were argued to be (post-hoc) evidence for implicit knowledge measures, but the hypotheses on these three criteria were not supported in Ellis' (2005) study (but see Gutiérrez, 2013, for evidence of systematicity). Systematicity and certainty have been predicted to be higher in implicit knowledge measures, but L2 learners could be more confident and respond more systematically when using their solid explicit and metalinguistic knowledge than when using their developing (or underdeveloped) implicit knowledge if they believe that they knew the correct grammar rules. For learnability, the acquisition of certain structures may be constrained by an individual's

age of acquisition (i.e., only learned implicitly), but we do not have enough data on the acquirability of many of the grammatical structures that would apply to all L2 learners.

Finally, the crucial factor that differentiated implicit knowledge and explicit knowledge measures was time pressure in R. Ellis (2005), Bowles (2011), and Zhang (2014). This factor has also been found to influence what types of knowledge are tapped into in other L2 studies (Erlam, 2006; Granena, 2013c; Gutiérrez, 2013; Y. Han & Ellis, 1998; Loewen, 2009), but all these researchers seem to assume that time pressure can limit the access to explicit knowledge enough to ensure that implicit knowledge is drawn upon. Time-pressure, however, does not guarantee the retrieval of implicit knowledge (DeKeyser, 2003, 2009). More efforts should be made to devise more fine-grained measures for both types of knowledge. Research on two types of measures that utilize time-pressured designs, GJT and EI, will be reviewed next. Critical examination of these two measures suggests that they are too coarse to reliably assess implicit knowledge and that they should be considered to be measures of automatized explicit knowledge.

2.2 Elicited Imitation

EI has been extensively used in first language (L1) acquisition research (e.g., Fraser, Bellugi, & Brown, 1963; Slobin & Welsh, 1973, Gallimore & Tharp, 1981), and it has also been applied to L2 populations (Bley-Vroman & Chaudron, 1994; R. Ellis, 2005, 2006; Erlam, 2006; Granena, 2010; Jessop, Suzuki, & Tomita, 2007; Lukyanchenko, 2011; Naiman, 1974; Tomita, Suzuki, & Jessop, 2009; Vinther, 2002). In general, EI involves the following three cognitive processes: (a) processing a spoken stimulus sentence; (b) reconstructing it with one's own grammar; and (c) reproducing it (Jessop, et al., 2007).

There is a consensus among L2 researchers that EI taps into linguistic knowledge (Bley-Vroman & Chaudron, 1994). In other words, EI involves reconstructive processing using internal

resources rather than parroting the stimulus sentences. It is, however, controversial as to exactly what resources it really taps into. .Elicited Imitation (EI) has been claimed to be the best measure for implicit knowledge (R. Ellis, 2009), and empirical evidence for EI as an implicit knowledge measure comes from the relationship of EI with other time-pressured tasks (i.e., a timed GJT and an oral narrative task). Ellis (2005), Bowles (2011), and Zhang (2014) are representative examples of this approach; they found that EI loaded most heavily on the implicit knowledge factor. Erlam (2006), who used the same data set from Ellis (2005), further examined whether EI is related to other time-pressured tests (i.e., an oral narrative task and the listening and speaking subtests of the IELTS test). Results showed that the time-pressured tests were highly correlated with the EI; time pressure thus appears to be the strongest factor underlying the performance on the group of tests that includes the EI, and is assumed to influence the awareness of language use.

As discussed above, the present study makes a distinction between explicit and implicit based on awareness in the use of linguistic knowledge; we argue that it is superior to the time pressure criterion because even when the task is done under time constraints, L2 learners may still be capable of using linguistic knowledge *with awareness* (DeKeyser, 2003, 2009). Access to explicit knowledge involves linguistic use *with* awareness even if the execution is rapid or automatic (i.e., automatized explicit knowledge), which is distinguished from the use of linguistic knowledge *without* awareness (i.e., implicit knowledge). This operationalization may imply that *fully* automatized explicit knowledge does not require awareness at all either (DeKeyser, 2003). If this were the case, it would be nearly impossible behaviorally to tease apart implicit knowledge from *fully* automatized explicit knowledge in the awareness criterion. A recent study, however, provides evidence that it is possible to devise linguistic tasks that can

distinguish automatized explicit knowledge from implicit knowledge. Note that the declarative/procedural and explicit/implicit distinctions overlap but are not equivalent (DeKeyser, 2009). The explicit/implicit distinction is essentially made by awareness, whereas the declarative/procedural distinction refers to whether it can be used for cognitive/psychomotor skills or not. There are certain cases where explicit/implicit is dissociated from the declarative/procedural distinction. Procedural (even automatized) knowledge is not necessarily implicit knowledge because it can result from explicit declarative knowledge; hence, automatized explicit knowledge can stand as a distinct construct from implicit knowledge. Additionally, implicit knowledge is not necessarily procedural, either. For instance, intuitive knowledge such as category prototypes and chunk information can be non-procedural implicit knowledge. The current study only focuses on tapping into proceduralized (and partially automatized) representations; both explicit and implicit knowledge are accessed quickly, but they are distinguished based on the awareness criterion.

Suzuki and DeKeyser (in press) hypothesized that two different types of cognitive processing might be at work during the EI task: processing of auditory stimuli and production (imitation). In the processing stage of EI, since an auditory stimulus disappears quickly, the listeners do not know whether/when errors will occur, and their attention is directed to meaning. Therefore, they have virtually no chance to deploy linguistic knowledge intentionally or consciously. Only if participants possessed implicit knowledge, they would be able to register[4] grammatical errors in the processing component of EI. In the production stage, in contrast, L2 speakers might be able to use automatized explicit knowledge to imitate the sentence even under

[4] The word *register* is used in a technical sense, meaning detection *without awareness*, which is distinguished from detection within focal attention accompanied by awareness (conscious perception or noticing) (Schmidt, 2001).

time pressure, because there is still some room for them to pay attention to specific grammatical structures before/while repeating the sentence.

Given these assumptions, they compared the ability to register grammatical errors and that to repeat the sentence. In order to examine whether participants register the error during initial sentence processing, a psycholinguistic task called the word-monitoring task (Marslen-Wilson & Tyler, 1980) was incorporated within the EI. The word-monitoring task typically includes a target word embedded in an auditory sentence, and participants need to respond by pressing a button as soon as they hear it. The rationale of the task is that test-takers slow down to respond to a target word that appears after a grammatical error, which reflects the sensitivity to the errors. For instance, the response time to the monitoring word (i.e., *by*) in a sentence like "The book is being closely picked/*pick by the large group of curious students" will be delayed when the monitoring word appears after the ungrammatical part of the sentence. The reaction time differences between grammatical sentences and ungrammatical sentences index the registration of errors in real-time sentence processing (Grammatical Sensitivity Index: GSI); a larger GSI indicates more developed state of implicit knowledge.

Sixty-three advanced L2 Japanese speakers were administered with the EI and word-monitoring in a single task along with a metalinguistic knowledge test (a measure of explicit knowledge) and the Serial-Reaction Time (SRT) task (a measure of implicit learning aptitude). Results revealed no significant relationship between the EI score (imitation) and the SRT score. In contrast, the GSI (word-monitoring) was related to the SRT score *only* among the long LOR speakers. In contrast, metalinguistic knowledge was a significant predictor of the EI, whereas metalinguistic knowledge did not have any influence on the GSI. These results suggest that a word-monitoring task is a measure of implicit knowledge, whereas EI draws on more explicit

types of linguistic knowledge. Even though time pressure was imposed for imitation, it appears that participants were able to monitor their utterances before or/and while they imitated the sentence. This finding goes against the claims made by Ellis (2005) and others in the sense that EI may be a measure for *automatized explicit knowledge*, not implicit knowledge. Not only production tests like EI, but also processing tasks like GJTs, even administered under time pressure, can be contaminated by the use of automatized explicit knowledge.

2.3 Grammaticality Judgment Test

GJT has been widely used in L1 and L2 research, and all or some of the following three steps are typically included in the procedure: (1) identification of ungrammatical sentences, (2) correction of the errors, and (3) explanation of rules (R. Ellis, 2004). The third step always requires conscious metalinguistic knowledge to perform (i.e., explicit knowledge). It is much harder to claim what kind of knowledge is deployed in steps (1) and (2). From the perspective of the present study's purpose, the question is whether they draw on explicit or implicit knowledge, or a mixture of both. Previous SLA research has identified two key factors that influence the types of linguistic knowledge tapped by GJT. The primary factors reported in the literature are 1) time allowed in response in GJT and 2) types of items in GJT (grammatical sentences and ungrammatical sentences).

As discussed in R. Ellis (2005), Bowles (2011) and Zhang (2014), accumulating evidence indicates that imposing time pressure on GJT responses influences the types of knowledge that test takers use (Bialystok, 1979; Granena, 2013c; Gutiérrez, 2013; Y. Han & Ellis, 1998; Loewen, 2009). The rationale behind the time limitation is that three mental operations are involved in the GJT performance: 1) semantic processing (i.e., understanding the meaning of a sentence), 2) noticing (i.e., searching to establish whether something is formally incorrect in the sentence), and

3) reflecting (i.e., considering what is incorrect about the sentence and, possibly, why it is incorrect) (R. Ellis, 2004). R. Ellis (2004) argued that when the time is only allowed for the processing of (1) and (2), the GJT *tends to* draw on implicit knowledge, whereas an untimed test allows participants to go through all three processing operations, and therefore, draws on explicit knowledge.

The biggest challenge, on this assumption, is to give L2 speakers enough time to perform the first two steps, but not for the third step. If advanced L2 speakers can parse the sentence really fast, then they have some time left to reflect on their response. As shown in Table 3, some previous studies imposed a limit of 3 seconds (Bialystok, 1979) or 3.5 seconds (Y. Han & Ellis, 1998) on all the test items. This arbitrary cut-off point for all the items is not justified, because individual sentences differ in length, structural complexity, and word difficulty. More effort was made to calibrate the time limitation for each of the individual sentences (Bowles, 2011; R. Ellis, 2005; Granena, 2012). R. Ellis (2005) determined the length of response time on each test item by calculating the average response time of NSs' responses plus 20%. This method is still left with some arbitrariness, however. Another problem also arises when a GJT is timed; that is, test takers sometimes fail to respond to the test items within the given time. Loewen (2009), who analyzed the same data from R. Ellis (2005), reported that lack of response due to the time limit occurred for 11% and 19% of the test items in L1 and L2 speakers, respectively. In Granena's study, which employed exactly the same procedure as R. Ellis, similar proportions of no responses were obtained. In Gutierrez's study, the proportion of missing data was smaller due to the fact that the time limit was not based on the 20% rule; the researcher determined the time limit for each of the sentences roughly based on previous studies (i.e., 6-9 seconds). No solid criterion for time limitation can be determined due to the small number of studies and the

arbitrary cut-off points (e.g., NSs' average processing speed plus 20%). At the same time, missing data in the measurements, due to no response, threaten reliability and validity of measurements. In order to avoid no response items, a *speeded* GJT can also be designed, in which test takers are asked to make a grammaticality judgment as quickly as they can. The drawback of this approach is that less time pressure can be imposed on their response, and it can limit the access to explicit knowledge to a lesser extent. A more ideal design for a timed GJT may be to set the time limit on every individual test taker by setting the time limit for each test item based on that individual's processing speed for grammatical sentences. This could be one way of making efforts to assure that L2 speakers are given enough time to perform semantic processing and noticing, but not enough time for reflection. Still, this method is left with other issues such as how to calibrate the individual's sentence processing speed for a variety of structures.

Table 3. *Design of Timed GJTs in Previous Studies*

	Modality	Medium	Time Pressure	Ratio of No response
Bialystok (1979)	Aural	paper	3 seconds were given after the whole sentence was read	Not reported
Han & Ellis (1998)	Written	Computer	Sentences were presented for 3.5 seconds	Not reported
R. Ellis (2005)	Written	Computer	20%+NS average RT	11%(NS) 19%(L2)
Bowles (2011)	Written	Computer	20%+NS average RT	Not reported
Granena (2012)	Aural Written	Computer	20%+NS average RT	10.61%(auditory) 15.67%(written)
Gutierrez (2013)	Written	paper-based	Sentences were presented for 6-9 seconds, and 3 seconds were given for writing the answer	7.33%
Zhang (2014)	Written	Computer	20%+NS average RT	Not reported

Some researchers pointed out that grammaticality of test items also influences the constructs tapped into by GJT (Granena, 2013c; Gutiérrez, 2013; Loewen, 2009). Loewen (2009) found that there was no difference between scores of the grammatical and the ungrammatical items on the *untimed* GJT, because test takers had enough time for all three steps (semantic processing, noticing, and reflection). In contrast, the ungrammatical items were significantly more difficult than the grammatical items on the *timed* GJT because, according to Loewen, test takers did not have enough time to reflect (step 3) to reject ungrammatical sentences, and their scores were lower on the ungrammatical items.

Loewen's (2009) analyses pointed to a potential difference in test-taking processes between grammatical and ungrammatical items, and the same effect of time pressure seemed to explain this; grammatical sentences may require only steps (1) and (2) (semantic processing and noticing), whereas rejection of ungrammatical sentences requires the third step, reflection. Gutierrez (2013) set out to formally examine this assumption by asking whether grammatical and ungrammatical items load on two different factors (explicit and implicit knowledge). An untimed GJT, a timed GJT, and a metalinguistic knowledge test were administered to L2 Spanish speakers, and the GJT scores were calculated for grammatical items and ungrammatical items separately. These four variables were submitted to a confirmatory factor analysis along with the metalinguistic knowledge score. Two hypothesized models were compared: The first model (Grammatical/Ungrammatical) hypothesized that grammatical items of the timed and untimed GJTs would load on one-factor (implicit knowledge) and that the ungrammatical sections of both tests and the metalinguistic knowledge test would load on another (explicit knowledge). The second model (Timed/Untimed) hypothesized that both grammatical and ungrammatical items in the timed GJT would load on an implicit factor, and that both types of items in the untimed GJT

and the metalinguistic knowledge test would load on an explicit factor. Results showed that the grammatical/ungrammatical model was a better fit for the data, suggesting that responses to the grammatical items reflect implicit knowledge, while responses to the ungrammatical items reflect explicit knowledge. Although the difference of the scores on the timed and untimed GJT was also significant, which is consistent with the previous studies, the difference was larger for the comparison between grammatical and ungrammatical items than that for the comparison of the timed GJT and the untimed GJT.[5] Gutierrez thus argued that, based on the three processing stages of GJT proposed by Ellis, "learners are able to determine whether or not sentences are grammatical on the basis of their implicit knowledge, it thus follows that semantic processing and noticing can be carried out on the basis of this type of knowledge. Additionally, learners need to resort to their explicit knowledge to engage in reflecting." (pp. 6-7)

From a slightly different angle, Granena (2013c) also offered a piece of evidence showing the difference between grammatical and ungrammatical test items on GJT. She administered a timed aural GJT and an untimed written GJT and investigated the relationship with explicit learning aptitude as measured with the LLAMA test (Meara, 2005). Results showed that the explicit learning aptitude was related to the ungrammatical items, not to the grammatical items, only on the untimed written GJT. Explicit learning aptitude seems to be required to attain the ability to correctly reject the ungrammatical sentences, suggesting that ungrammatical items draw more on explicit knowledge.

As shown above, several SLA researchers have attempted to develop GJTs that draw on implicit knowledge by imposing time limitation or by examining the differential effects of

[5] The procedure of setting the time limit on responses in the timed GJT in Gutiérrez (2013) is different from the one in R. Ellis (2005) or Bowles (2011) (see Table 3); it might be the case that the reduced time pressure led to the better model fit of the grammatical and ungrammatical model in Guitierrez (2013).

grammatical and ungrammatical items. These two parameters influence their test takers' mental processes; thus, a different source of knowledge is employed. However, the separate use of grammatical and ungrammatical items cannot be justified. The biggest concern is that it is not clear what linguistic knowledge (of which target structures) is measured by the grammatical items. There are two types of cases where interpretations of the scores on the grammatical items become problematic. First, when a grammatical item is rejected, it is not clear what structure is unknown. Second, when test-takers who accept the grammatical items also fail to reject the corresponding ungrammatical items, it is not possible to interpret the score of grammatical items either.

At first sight, the factor of time pressure on the GJT seems to offer a better ground for assessing implicit knowledge; the biggest caveat, however, is that the GJT is inherently a task that asks participants to make a judgment on grammaticality, i.e., focus on form. Paying attention to form inevitably raises awareness of one's linguistic knowledge. This nature of the GJT should be taken as a major threat to its validity as an implicit knowledge measure, because it is still possible to access explicit knowledge even under time pressure when explicit knowledge is automatized (DeKeyser, 2003, 2009). The same criticism of the validity of EI can be made for the validity of the GJT (Suzuki & DeKeyser, in press): timed GJTs may be a measure for automatized explicit knowledge rather than implicit knowledge.

2.4 Problems in Existing Measures and Alternative Operationalization for Implicit Knowledge

The critique of the two measures in the previous sections indicates that the previous validation studies primarily rested on the assumption that time-pressure makes the tasks more conducive to the elicitation of implicit knowledge. The accumulating evidence suggests that

timed tests appear to draw on different types of knowledge; however, the caveat is that both measures leave much room, particularly for L2 speakers who possess automatized explicit knowledge, to have recourse to explicit knowledge. In addition to the problem of the time feature of tasks, conversion of ungrammatical sentences to grammatical sentences in EI seems to be conducted consciously in some L2 speakers (Chrabaszcz & Jiang, 2014; Suzuki & DeKeyser, in press). Existing measures are too coarse a measure for assessing implicit knowledge, and they may be better operationalized as automatized explicit knowledge measures.

The most utilized criterion for implicit knowledge, time pressure, cannot always shut off the access to automatized explicit knowledge. Following Suzuki and DeKeyser (in press), the current study focuses on the criterion of awareness and proposes an alternative operationalization for implicit knowledge. Examining the awareness of use is often hard, and subjective retrospective reports on awareness are not always reliable.[6] The most promising and objective methodology is to employ an online psycholinguistic technique like the word-monitoring task. When measuring real-time sentence processing while speakers' attention is directed to meaning, they have much fewer opportunities to apply linguistic knowledge consciously because the processing occurs so fast and automatically. Only implicit knowledge makes it possible to operate within this short period of time (i.e., hundreds of milliseconds) while sentence processing is focused on meaning. They thus operationalize implicit knowledge as registration (in the restricted sense) of errors during real-time sentence processing for comprehension. The registration of specific grammatical features in linguistic input appears to be closely tied to the

[6] In the implicit *learning* paradigm in research in the field of cognitive psychology and SLA, retrospective verbal report and subjective report are often used to measure whether the product of learning is unconscious or conscious (Dienes, 2004, 2007; Dienes & Scott, 2005; Rebuschat, 2013; Rebuschat, Hamrick, Sachs, Riestenberg, & Ziegler, 2013). The present study focuses on the development of objective measures rather than subjective measures.

real-time prediction in sentence processing. In the word-monitoring task, the delayed response in ungrammatical sentences compared to the corresponding grammatical sentences (i.e., an index of registration) results from two sources: 1) the facilitation to respond to the target word in grammatical sentences that L2 speakers predict based on the linguistic information (e.g., case markers) and 2) the delay to the target word in ungrammatical sentences that conflicts with the one that they predict from the ungrammatical linguistic input (e.g., wrong case information).[7] We thus propose that prediction is the key factor that makes the registration occur; we take the tentative view that registration of specific (un)grammatical structures is largely driven by predictive sentence processing. By postulating registration and prediction based on specific grammatical structures as related constructs, we can extend options of online psycholinguistic methods from reaction-time tasks to eye-tracking task. In the next section, we will discuss reaction-time measures with focus on a word-monitoring task and a self-paced reading task. After that, we introduce a still newer method in the L2 field, an eye-tracking while-listening task (i.e., visual-world task).

2.5 Reaction Time Measures

Over the decades an increasing interest in a psycholinguistic approach to SLA has developed, leading to an every increasing use of reaction time to examine online sentence processing in L2 (see Jiang, 2011 for review). Many SLA researchers have successfully applied psycholinguistic techniques used for L1 populations to examine L2 sentence processing. Representative tasks include the self-paced reading task (Foote, 2011; Jiang, 2004, 2007; Jiang,

[7] Not all grammatical structures generate prediction in the same way. For instance, in an English sentence that contains violation in the third person *s* (e.g., Emily often go shopping on weekends), the delay in responding to the word right after the grammatical error (i.e., shopping) is not caused by prediction. The current study focuses on Japanese grammatical structures that involve predictions, which will be explained later.

Novokshanova, Masuda, & Wang, 2011; Juffs, 1996; Juffs & Harrington, 1995; Roberts & Liszka, 2013), the adapted serial reaction time task, inspired by the contextual cuing paradigm and the derived attention paradigm (Leung & Williams, 2011, 2012), the word-monitoring task (Granena, 2012, 2013b; Jiang, Hu, Lukyanchenko, & Cao, 2010; Suzuki & DeKeyser, in press), and the sentence-picture matching task (Jiang, 2011). The common logic behind these tasks, as explained in the previous section, is that the reaction time difference between grammatical sentences and ungrammatical sentences indicates the online registration or sensitivity to grammatical errors. The advantage of using these types of reaction time methods, over form-focused GJTs, is that we can indirectly measure their grammatical sensitivity without asking them to make grammaticality judgments. This is because reaction time is measured while participants are paying attention to meaning rather than form, because of the inclusion of comprehension questions. One of the most frequently used methods is a self-paced reading task. The L2 self-paced reading studies are briefly introduced first, and then a similar technique, the word-monitoring task, is taken up to discuss its advantages and disadvantages over the self-paced reading task.

One of the earlier studies that applied the self-paced reading task to investigate acquisition of L2 morphosyntactic structures was conducted by Jiang (2004). He examined whether L2 English speakers with L1 Chinese were sensitive to the errors about plural markers. Participants read sentences like (1) and (2) by pressing keyboard buttons to show the next word, and were asked to answer comprehension questions after the sentences in half of the trials. The question was whether L2 speakers can show sensitivity to grammatical errors, which is reflected by the slow-down of reading time at the exact word and subsequent word where ungrammaticality occurs in sentence 2 (i.e., “was” and “about”).

1. The bridge to the island was about ten miles away.
2. *The bridge**s** to the island was about ten miles away.

Although NSs showed the expected pattern of reaction time, L2 speakers did not show any sensitivity. This morphological insensitivity measured with the online task has been replicated in follow-up studies with English L2 speakers (Jiang, 2007; Jiang, et al., 2011), and other structures have been investigated such as Spanish gender agreement structures (e.g., Foote, 2011). These studies have demonstrated the validity of the self-paced reading task.

Online sensitivity to grammatical errors, as investigated by Jiang and other researchers, is usually associated with the question of whether second language learners can acquire *automatized* or *integrated* linguistic knowledge of morphological marking. *Integrated knowledge* and its processing counterpart (i.e., *automatic competence*) are defined as linguistic knowledge that enables the use of linguistic knowledge spontaneously in both the productive and receptive use of language (Jiang, 2007). Their operationalization could be translated to the measurement of implicit knowledge; linguistic knowledge that can be used spontaneously in real-time sentence processing without awareness indicates implicit knowledge.

As described in the previous section, the word-monitoring technique is also similar to the self-paced reading task. It has recently started to be utilized by SLA researchers. The target structures tested included English tense-marking (Jiang, et al., 2010), Spanish agreement structures and non-agreement structures (Granena, 2013b), and Japanese particles (Suzuki & DeKeyser, in press). These studies form the foundation for the use of the word-monitoring task as a valid assessment tool for implicit knowledge.

The task had advantages over the self-paced reading task because it: 1) is more likely to prevent L2 learners from using explicit knowledge, particularly in the aural modality, 2) requires participants to engage in the dual task of monitoring the word and comprehending the sentence, which also demands the use of automatized knowledge, 3) is potentially more appropriate for a broader range of L2 populations, such as those who may lack literacy in the target language (e.g., heritage speakers), and 4) reduces the tediousness of self-paced reading tasks, in which participants are required to press buttons repeatedly to read sentences[8] (Jiang, 2011). In sum, both self-paced reading and word-monitoring tasks have been utilized successfully and will be used in the present study.

2.6 Eye-Movement Measures

Reaction-time based research has stood as a gold standard for psycholinguistic studies, but a more fine-grained measurement technique has been developed and widely used in L1 research—eye-tracking. The focus of this section will be on an eye-tracking technique that is used while the participant is listening— the visual world paradigm.[9] As Sedivy (2010) summarized, eye-movements are useful to reveal language processing because 1) people tend to direct their eye gaze to things they are attending to in their visual environments, 2) eye movements are generated by linguistic input, and 3) eye movements reflect highly incremental linguistic interpretation. Research to date has investigated real-time sentence processing with the visual-world paradigm with L1 adult and children populations, but this method has recently been extended to L2 populations. As a demonstration of this relatively new technique to the SLA field,

[8] The self-paced reading task allows for the observation of unfolding sentence processing at each point of the sentence (i.e., word-by-word), which offers more data points than a word-monitoring task.

[9] Eye-tracking while reading has also been used in L2 research (e.g., Keating, 2009; Foucart & Frenck-Mestre, 2012), but this is beyond the focus of the present study.

a seminal study is introduced first, and then L2 research will be presented with the comparison of L1 research.

2.6.1 The Visual-Word Paradigm in L1 Research. The visual-world paradigm was first devised to investigate the real-time processing of ambiguous sentences (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). The technique has been applied to different types of linguistic structures in different languages, and it has revealed that native speakers incrementally process linguistic input very rapidly, taking into account the visual context (see Tanenhaus & Brown-Schmidt, 2008; Huettig, Rommers, Meyer, 2011 for review).

A seminal study by Tanenhaus et al. (1995) investigated whether visual context facilitates the resolution of ambiguous prepositional phrases. English native speakers were presented aurally with an ambiguous sentence like "Put the apple on the towel in the box." The first prepositional phrase (i.e., on the towel) is temporarily ambiguous as to whether it modifies the noun (apple) or it denotes the goal location of the verb (put). Two types of visual arrays were created (Figure 1): the critical manipulation was that the left visual display had only one apple on a towel (the one-referent condition), whereas the right visual context included two apples, one on a towel and one on a napkin (the two-referent condition). The two-referent display was predicted to facilitate the disambiguation of the target sentence because the parser was made more likely to process the prepositional phrase as a modifier of the noun. This prediction was supported by the results that fewer proportions of looks were directed to the incorrect goal (i.e., empty towel) in the two-referent condition than in the one-referent condition. This seminal work demonstrated that "eye movements can be used to observe under natural conditions the rapid mental processes that underlie spoken language comprehension" (p. 1634). This finding has been replicated extensively in the L1 literature (Ferreira, Foucart, & Engelhardt, 2013; January, Trueswell, &

Thompson-Schill, 2009; Novick, Thompson-Schill, & Trueswell, 2008; Tanenhaus, Chambers, & Hanna, 2004), which confirms the validity of this paradigm.

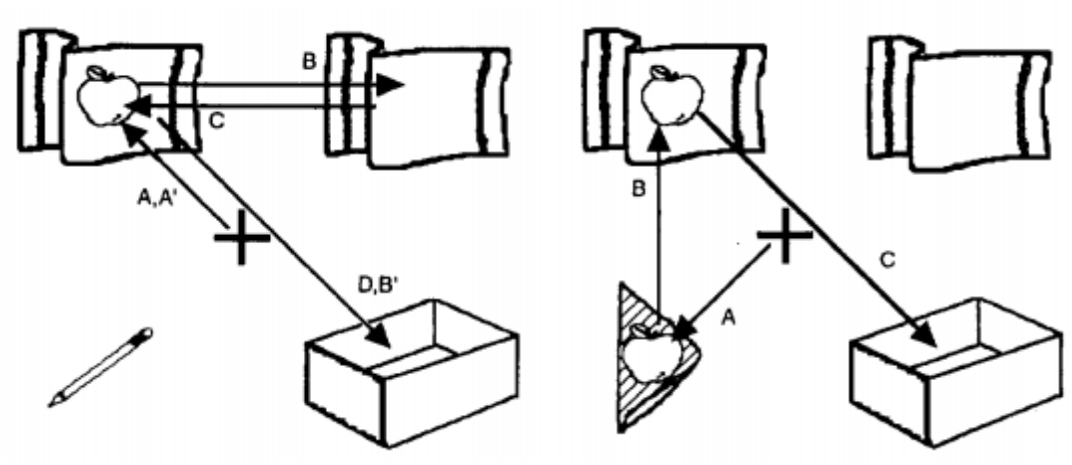


Figure 1. Visual Display in Tanenhaus et al. (1995)

Visual-world paradigms have been extended to test various linguistic features across different languages: English articles (Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002), English quantifiers (Huang & Snedeker, 2009), the English tense system (Altmann & Kamide, 2007), English pronouns and reflexives (Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000; Elsi Kaiser, Runner, Sussman, & Tanenhaus, 2009; Runner, Sussman, & Tanenhaus, 2003), Finnish pronouns and demonstratives (E. Kaiser & Trueswell, 2008), sub-categorization information of English verbs (Altmann & Kamide, 1999), Japanese case markers (Kamide, Altmann, & Haywood, 2003), German case markers (Kamide, Scheepers, & Altmann, 2003), Spanish gender marking (Lew-Williams & Fernald, 2007, 2010), and Chinese classifiers (Huettig, Chen, Bowerman, & Majid, 2010). Some recent research has started to apply this method to L2 speakers, and the technique has proven to be advantageous in directly assessing linguistic knowledge without contamination from explicit knowledge. This L2 research will be reviewed next.

2.6.2 Application of the Visual-World Paradigm to L2 research. L2 visual-world studies have recently started to investigate whether L2 speakers show the same incremental processing as L1 speakers by using an identical or an adapted setup of experiments in the L1 literature (Ellert, 2013; Grüter, Lew-Williams, & Fernald, 2012; Hopp, 2013; Kapnoula, Packard, Gupta, & McMurray, 2015; Lew-Williams & Fernald, 2007, 2010; Trenkic, Mirkovic, & Altmann, 2014). These studies attempted to compare L1 and L2 speaker processing of the same structure. Before presenting the L2 research inspired by each of these studies, we will briefly describe the corresponding L2 study first. The original L1 studies tested three different structures: Spanish gender marking systems (Lew-Williams & Fernald, 2007), the English article system (Chambers, et al., 2002), and the Japanese case-marking system (Kamide, Altmann, et al., 2003). The L1 studies will be described first to lay the ground for L2 research to be presented subsequently.

Acquisition of Spanish gender markings has been extensively investigated with a variety of research methods. Lew-Williams & Fernald (2007) is the first study that applied the visual world paradigm to this issue. They tested Spanish L1 adults and young children (34-42 months old) to examine whether they could use the grammatical gender information rapidly. Participants were presented aurally with target sentences in which the definite article and the noun were embedded in the carrier sentence, such as "Encuentra la pelota. ¿La ves?" (Find the ball. Do you see it?) or "¿Dónde está la vaca? ¿La ves?" (Where is the cow? Do you see it?). The critical manipulation was done to the visual context comprised of two objects: the names of the pictures were either both masculine or both feminine in the same-gender condition, whereas in the different-gender condition, the two nouns were different in grammatical gender. It was expected that the gender-marking article would facilitate anticipating the subsequent noun when they

heard it only in the different-gender condition. Results supported the prediction, meaning that both adult L1 speakers and young children were able to exploit the grammatical gender cue in real time to predict the forthcoming noun.

All three subsequent L2 studies were motivated by Lew-Williams and Fernald (2007) and aimed at extending the experiment to L2 populations (Dussias, Valdés Kroff, Guzzardo Tamargo, & Gerfen, 2013; Grüter, et al., 2012; Lew-Williams & Fernald, 2010).

With the identical experimental set up as Lew-Williams and Fernald (2007), Lew-Williams and Fernald (2010) tested L2 Spanish classroom learners with English as an L1 (they had received approximately five years of classroom instruction). Contrary to the L1 research, the results showed that L2 speakers were not able to utilize grammatical gender marking to predict the noun in real-time sentence processing. In a subsequent study, 19 more proficient L2 Spanish speakers were recruited and tested (Grüter, et al., 2012). They had obtained an overall score within the top two tiers of a standardized speaking test (the Versant Spanish Test), and 13 out of the 19 speakers had been using Spanish in their jobs (translator, interpreter, or language teacher). On an off-line test of receptive knowledge of gender agreement, L2 speakers showed comparable performance to L1 speakers. Online results, however, did not change from the previous study: L2 speakers were still less adept at using grammatical gender cues to facilitate the subsequent processing.

Given the findings from the two studies, Grüter et al. (2012) suggested that the strength of association between nouns and gender nodes in the mental lexicon is weaker in L2 speakers, which reflects the slower and more effortful retrieval of gender information in real-time L2 use. Grüter et al.'s (2012) interpretation mainly focused on strength of association between the nouns

and gender, but the findings can also be framed against the explicit knowledge (off-line performance) versus implicit knowledge (real-time processing) distinction.

Another follow-up study of Lew-Williams and Fernald (2007) was conducted by Dussias et al. (2013). They investigated how L2 proficiency and the existence of a gender-marking system in L1 would modulate the real-time processing ability. They recruited 18 English-speaking learners of Spanish from a large United States institution and 16 Italian learners of Spanish who were completing a year of university study in Granada. English-speaking Spanish learners were further divided into higher and lower proficiency groups based on a standardized test of Spanish (i.e., Diploma de Español como Lengua Extranjera "Diploma of Spanish as a Foreign Language"; DELE). An adapted version of the visual-world task in Lew-Williams and Fernald (2007) was administered in this study. The carrier sentences were more complex than in the Lew-Williams and Fernald (2007) study: half of the critical noun phrases appeared in the middle of the sentence, e.g., for *el reloj* "the clock:" *el estudiante estaba dibujando el reloj que vio ayer* (the student was drawing the clock that he saw yesterday) and the other half at the end, e.g., *el niño miraba a su hermano mientras fotografiaba el reloj* (the boy watched his brother taking a picture of the clock). In addition, participants were required to make a semantic plausibility judgment after each trial to ensure that they focused on meaning. Results showed that more proficient L2 speakers with L1 English showed the same pattern as NSs, whereas the less proficient L2 group with L1 English did not show the anticipatory eye-movement to the target noun. Although proficiency of L2 speakers with L1 Italian was even lower than the less-proficient group of L1 English speakers, they were able to predict the noun by using the feminine gender marking information (not the masculine gender marking). As Dussias et al. (2013) put it, their "study is the first to provide empirical evidence demonstrating the rapid use of gender-

marked information in articles to speed up noun recognition when attention is directed to other features of processing (i.e., in-depth semantic processing)" (Dussias et al., 2013, p. 377). Incremental processing while attention is directed to other features than form is consistent with the interpretation that it indicates implicit knowledge.

Another target structure examined is post-nominal case markers in the head-final language Japanese (Kamide, Altmann, et al., 2003). In English, a verb drives the predictive processing (Kamide & Altmann, 1999 and experiments 1 and 2 in Kamide & Altmann, 2003). It has been demonstrated that when English L1 speakers listen to the sentences (3) and (4) in the context of a visual scene (consisting of a woman, a man, butter, and bread), they generate anticipatory eye movements, at the verb position before hearing the Goal, towards the bread in (3) (the bread is a plausible Goal) and towards the man in (4) (the man is a plausible Goal).

3. The woman will spread the butter on the bread.
4. The woman will slide the butter to the man.

Given that the verb always appears at the end of the sentence in Japanese, Kamide and Altmann (2003) hypothesized that predictive processing should be driven by (pre-verbal) case-markers in Japanese. Experiment 3 in Kamide and Altmann (2003) examined whether case-marking particles would allow Japanese L1 speakers to anticipate the subsequent arguments without hearing a verb. Participants' eye movements were tracked while listening to the following two sentences in the visual scene in Figure 2.

5. Dative condition.

Weitoresu-ga kyaku-ni tanosigeni hanbaagaa-o hakobu.

waitress-NOM customer-DAT merrily hamburger-ACC bring.

The waitress (will) merrily bring the hamburger to the customer.

6. Accusative condition.

Weitoresu-ga kyaku-o tanosigeni karakau.

waitress-NOM customer-ACC merrily tease.

The waitress (will) merrily tease the customer.

Figure 2. Visual Scene in Experiment 3 in Kamide and Altmann (2003)

After hearing waitress-NOM customer-DAT in the dative condition (5), Kamide and Altmann (2003) predicted that eye movement would be directed to "hamburger," which could be plausibly transferred by the waitress to the customer. In contrast, fewer anticipatory looks towards the hamburger were expected in the accusative condition than in the dative condition, because the fragments of NP-NOM and NP-ACC do not usually anticipate any more arguments. This is exactly what they found, suggesting that syntactic information of case-markings allows the parser to predict subsequent argument structures in real time.

Based on this finding in L1, Mitsugi and MacWhinney extended the experiment to Japanese L2 speakers (Kapnoula, et al., 2015). Twenty-seven classroom L2 learners (450–600 hours of formal classroom), as well as L1 control groups, were tested with the adaptive task from Kamide and Altmann's (2003) experiment. The results replicated the previous findings and further showed that L2 speakers' eye movements did not anticipate a plausible theme object incrementally.

Another line of investigation focuses on the acquisition of English articles. The visual world paradigm experiment in Chambers et al. (2002) examined whether L1 English speakers can use the definite/indefinite information of English determiners. Definiteness refers to whether a referent is uniquely identifiable (definite) or neutral to uniqueness (indefinite) (Lyons, 1999). For instance, a definite description like "Give me the book" identifies a unique referent to listeners, whereas an indefinite description like "Give me a book" implies that there is more than one book in the situation.

Trenkic et al. (2014) adapted the visual-world paradigm from Chambers et al. (2002) to examine whether L2 English speakers with L1 Chinese who recently arrived in the United Kingdom (the median length of stay in an English-speaking country was two months). In the visual-world task, participants were presented with a clip-art picture on the computer display and heard a stimulus sentence while eye-movements were recorded. Participants heard an auditory sentence containing a definite or indefinite article and a location noun such as "The pirate will put the cube inside **the/a** can. While hearing the sentence, they saw the display consisting of one target location (one-referent trials) or two target locations (two-referent trials). Fixations to the target goal were analyzed in terms of definiteness of articles and the number of target locations. Despite the relatively small amount of L2 experience in the United Kingdom, L2 speakers

demonstrated sensitivity to the definiteness information in real time like native speakers. That is, in the one-referent trials, the eye-movements converged onto the goal faster when hearing the definite description than the indefinite description. In contrast, in the two-referent trials, the faster convergence to the goal was found when hearing the indefinite description than the definite description.

In summary, previous research that employed visual-world tasks has presented mixed findings with regard to predictive ability in L2. Some studies show that it is possible for L2 learners to show the native-like anticipatory eye-movement patterns, whereas others do not. Individual differences factors such as L2 proficiency and immersion experiences seem to moderate the attainment.

2.6.3 Advantages of Eye-tracking methods. Advantages of applying the visual-world paradigm to L2 research are summarized as follows: 1) a direct measure of fast and ballistic linguistic processing in real time, 2) higher ecological validity, 3) simple tasks are usually used, which can be applied to wider populations, and 4) no ungrammatical sentences are required.[10]

First, eye-movements are a direct measure of rapid language processing. One of the limitations in the reaction time measures is that reaction time is always mediated through the button responses. Anecdotally, some NSs are less disturbed by the grammatical errors in sentence processing, and this is partly because some people seem to be able to override the grammatical error to hit the button even though they notice the errors. Eye movement is automatically launched after 200ms to 250ms once the movement is programmed (Matin, Shao,

[10] A disadvantage of this paradigm may be practicality, because it requires expensive equipment or laborious coding and analysis if you use inexpensive equipment. A more practical and useful technique may be a mouse-tracking method (Freeman & Ambady, 2010; Freeman, Dale, & Farmer, 2011). Tanenhaus et al.'s (1995) finding has recently been replicated by this technique (Farmer, Anderson, & Spivey, 2007; Freeman, et al., 2011), but this is beyond the focus of the present dissertation.

& Boff, 1993), which can index a purer measure for language processing with less interference. Second, the visual-world paradigm enjoys higher ecological validity than any other psycholinguistic tasks because participants can be tested in a more naturalistic (yet well-controlled) context (see, for example, Brown-Schmidt, 2009, for the application of the visual-world paradigm in conversational interaction). Third, it can be applied to a wide range of populations, given the simplicity of the task required; it has been used successfully with pre-school populations (Huang & Snedeker, 2013; Lew-Williams & Fernald, 2007; Trueswell, Sekerina, Hill, & Logrip, 1999) and children with Specific Language Impairment (SLI) (Andreu, Sanz-Torrent, & Trueswell, 2013). These populations are harder to test by conventional methods, compared to healthy adult L1 speakers (e.g., with GJT), because their explicit judgments on grammaticality are harder to elicit. This concurs with the challenge for SLA researchers, from the opposite perspective, because L2 learners rely on explicit knowledge to a larger extent, which makes it harder to tap into implicit knowledge. The visual-world paradigm can draw on implicit knowledge without contamination from explicit knowledge. Finally, the visual-world task does not require inclusion of ungrammatical sentences, which contrasts with reaction-time-based methods like the word-monitoring task.[11] Including grammatical errors is likely to induce a higher awareness of linguistic surface forms, especially after encountering similar errors over the course of the experimental task. The visual-world paradigm is particularly advantageous, as there is no need to include ungrammatical sentences.

[11] The picture-sentence matching task, which is also one of the reaction-time-based methods, can measure grammatical sensitivity without including ungrammatical sentences, however (Jiang, 2011).

In sum, the advantages of the visual-world paradigm provide the purest index[12] of incremental sentence processing, which reflects implicit knowledge. An emerging line of research into incremental processing with the visual-world paradigm opens an avenue to tackle a core issue in the SLA field, which will be presented in the next section.

2.7 Interface Issues of Explicit and Implicit Knowledge and Learning

One of the biggest controversies in the SLA field, debated for decades, is the interface problem of explicit and implicit knowledge, namely, whether explicit knowledge leads to implicit knowledge and whether implicit knowledge can lead to explicit knowledge. The latter issue that implicit knowledge becomes explicit is much less controversial, at least for native speakers (Bialystok, 1994); however, SLA researchers take different positions on whether explicit knowledge leads to implicit knowledge in SLA.

The review of the literature suggests that SLA researchers primarily focus on the product of learning (i.e., linguistic knowledge), but the interface issue also pertains to learning processes. Knowledge and learning are related but distinct; the first set refers to the end-products of learning, whereas the second refers to the processes of learning (Schmidt, 1994b). Among various operationalizations of explicit and implicit learning is the one stipulated by the skill acquisition theory: that a deliberate conscious learning process can lead from declarative knowledge to proceduralized and automatized knowledge (DeKeyser, 2007b; Lyster & Sato, 2013; McLaughlin, 1987), whereas implicit learning refers to the learning without intention (incidental learning) and without awareness (Long, 2015; Williams, 2009) and the tallying of sequential dependencies and distributional properties (N. C. Ellis, 2002, 2005). How do these

[12] Although a growing body of neurolinguistics studies has contributed to the understanding of the issue (e.g., Kotz, 2009; Morgan-Short, 2014 for review), the present study focuses on behavioral experimental methods because the interpretation of components in neurolinguistics studies are yet less straightforward.

learning processes and the product of learning influence each other? It is usually the case that the positions to the interface issue are broadly categorized into three: the non-interface, the strong interface, and the weak interface (see Han & Finneran, 2013 for a recent review), which will be discussed in detail.

2.7.1 Non-Interface Position. The most well-known proponent of the non-interface position is Krashen, who distinguished *acquisition* from *learning*, in the early days of SLA (Krashen, 1981, 1982, 1985, 1994). According to him, "'acquisition' is a subconscious process identical in all important ways to the process children utilize in acquiring their first language, whereas 'learning' is a conscious process that results in 'knowing about' language" (Krashen, 1985, p. 1). This distinction, despite its vague definition of constructs like consciousness (cf. Schmidt, 1994a), corresponds to types of learning: explicit and implicit learning. The champion of the non-interface position postulated no interaction between the two types of knowledge, rejecting the idea that explicit knowledge can be converted into implicit knowledge. He further claimed that only comprehensible input can result in intake for L2 development and acknowledged no or only a marginal role for an explicit mode of learning (e.g., grammar instruction and error correction). According to him, learned knowledge or explicit knowledge can only be used as a "Monitor" under the restricted condition in which 1) enough time is available to think about and apply rules, 2) attention is focused on form or correctness, and 3) the correct grammatical rule is known by the L2 speaker (Krashen, 1981). With the marginal role given for explicit knowledge and learning, Krashen maintains that L2 implicit knowledge is acquired through an independent route from explicit knowledge; he does, however, acknowledge the potential impact of explicit knowledge on the development of implicit knowledge in a very

restricted sense: by facilitating communication at early stages and hence generating more comprehensible input.

Taking a connectionist approach, Hulstijn (2002) also embraced the non-interface position, with a more detailed description of explicit and implicit learning. According to Hulstijn, "implicit learning is an autonomous process, taking place whenever information is processed receptively (through hearing and seeing), be it intentionally and deliberately or unintentionally and incidentally. Implicit learning is not under conscious control" (p. 206). Hulstijn took a connectionist approach: any kind of L2 use (reading, listening, speaking, writing) automatically adjusts the connection weights in the network. In contrast, "explicit learning is the construction of explicit, verbalizable, metalinguistic knowledge in the form of symbols (concepts) and rules, specifying relationships between concepts" (p. 206). He maintained that explicit representation or knowledge is never transformed into implicit knowledge. More specifically, implicit learning takes place autonomously, beyond conscious control in L2 use, and the only choice left for L2 learners is whether or not to engage in the deliberative process of concept formation and concept linking (i.e., explicit learning). A critical difference from Krashen's theory is that Hulstijn valued the usefulness of explicit knowledge for language processing and production as a resource where and when implicit knowledge is not (yet) available. As he summarizes his position: "although explicit knowledge cannot be transformed into implicit knowledge neurophysiologically, explicit grammar instruction may indirectly be beneficial to the establishment of implicit knowledge" (Hulstijn, 2007, p. 701).

Further convincing support for the non-interface position comes from a neuropsychological perspective (Paradis, 1994, 2004, 2009). Paradis (1994) provided evidence supporting the dissociation of two distinct neuro-functional systems of implicit knowledge in

procedural memory and explicit knowledge in declarative memory. The non-interface position is put forward strongly in his argument that explicit and implicit knowledge do not share information; they do not exchange data; they do not interact. Similarly to Hulstijn and in contrast to Krashen, Paradis values the importance of explicit (metalinguistic) knowledge and claims that its effect is only indirect. An analogy given by Paradis (2009) was that if you read a book about a particular plant that cures a particular disease, and you are cured after taking the plant, you cannot say that the book interfaced with your being cured. The process of the cure was performed only internally, and the information on the book is not the cause of the cure (i.e., indirect effect). According to Paradis, both explicit knowledge and implicit knowledge continue to co-exist and develop via independent routes in parallel; what occurs in the course of development is switching from using explicit knowledge to implicit knowledge. His main claim was that it is never the case that explicit knowledge transforms into implicit knowledge; rather, the use of explicit knowledge can be replaced by the use of developed implicit knowledge.

2.7.2 Weak-Interface Position. The weak-interface position is proposed by R. Ellis (1990; 2008). It partially acknowledges the interface between the two types of knowledge depending on linguistic structures. According to R. Ellis, conversion of explicit knowledge into implicit knowledge is possible in the case of variational features (e.g., copula "be"), while the transformation from explicit to implicit knowledge is impossible in developmental features (e.g., negation). This position seems to be informed by the Processability Theory (Pienemann, 1998, 2005), showing that developmental features, which follow the fixed developmental trajectory, can only be facilitated by explicit knowledge only when L2 learners are developmentally ready. R. Ellis interpreted this as the effectiveness of explicit knowledge being limited to indirect effects on development of implicit knowledge of developmental features. R. Ellis (2009),

however, admitted that he recently conceded that indirect effects of explicit knowledge are more prevalent (see R. Ellis, 2006).

Another proponent of the weak interface position is N. Ellis (2002, 2005). He believes that explicit and implicit knowledge are distinct and dissociated; explicit knowledge does not turn into implicit knowledge. His position is similar to the non-interface position in that sense; however, he claimed that there is an interaction. According to N. Ellis (2002), the primary learning mechanism for L2 learning is implicit, tallying, priming, and strengthening through language use. Explicit learning plays a major role in the initial registration of linguistic patterns. Drawing on Schmidt's noticing hypothesis (Schmidt, 1990, 1993, 1994a, 1994b, 2001), which held that noticing is a necessary and sufficient condition for the conversion of input to intake in SLA, he regards conscious awareness as playing an important role in tuning of the input and the initial consolidation of a unitary representation. N. Ellis's position concurs with the non-interface position at some level, but he maintains that explicit and implicit knowledge interact *directly* as opposed to the non-interface position. He claims that the degree of influence of metalinguistic information on the nature of that processing is so profound that claims of interface and interaction seem fully justified. This direct interface position is based mainly on the grounds that the meta-analysis findings in Norris & Ortega, form-focused instruction has a medium-sized effect on free-constructed production measures, which is further confirmed by R. Ellis's review (2002). N. Ellis seems to interpret these findings as supporting evidence for the influence of explicit knowledge on implicit learning. However, free constructed production measures might not tap into implicit knowledge, and his conclusion should be taken cautiously (see more extensive evaluation of the direct-effect issue in Chapter 3 of Paradis, 2009).

2.7.3 Strong Interface Position. The strong interface position is often associated with skill acquisition theory (DeKeyser, 1995, 1997, 2007a, 2007b; Lyster & Sato, 2013; McLaughlin, 1987), which originated from Anderson's Adaptive Character of Thought (ACT) theory (Anderson, 1982, 1996, 2005). The skill acquisition theory stipulates two types of knowledge (declarative and procedural knowledge); second language learners first learn declarative knowledge (i.e., knowledge about the grammatical rules), and then engage in deliberate practice and develop procedural knowledge that allows them to use a second language faster and more effortlessly. The procedural knowledge can be automatized with further extensive practice, resulting in more accurate, fast, spontaneous, and effortless use. The theory postulates that declarative knowledge can lead to procedural and automatized knowledge through systematic and extensive practice. Highly automatized knowledge is often considered "functionally" equivalent to implicit knowledge in the narrow sense of without awareness (DeKeyser, 2003), but automatization of explicit knowledge is never equivalent to explicit knowledge turning into implicit knowledge. DeKeyser mentioned in his earliest studies that "no claim is being made here that the automatized knowledge documented in this study is equivalent to the implicit knowledge typically acquired in the native language" (DeKeyser, 1997, p. 215). More recently, he emphasized the importance of distinction between explicit and implicit knowledge from a purely psycholinguistic point of view, as explicit knowledge never becomes implicit knowledge, drawing on the work by Paradis (2009) (DeKeyser, in press).

In sum, skill acquisition theory focuses on the detailed documentation of how explicit learning leads to the knowledge that can be deployed more automatically. At the same time, it acknowledges that a different independent learning route (i.e., implicit learning) does exist and even predicts that late L2 learners who had massive input and practice for long periods of

residence in a foreign country may start to rely on implicit learning mechanisms rather than explicit learning (DeKeyser, 2007b).

2.7.4 Summary of Interface Positions. The debate on the interface issue reviewed above, i.e., whether explicit knowledge influences implicit knowledge, is based on evidence from research in cognitive psychology and neuroscience. It remains a conceptual and an empirical question whether explicit knowledge exerts influence on the development of implicit knowledge in SLA. Having closely inspected all the claims by representative researchers in the field, the distinction between the three interface positions does not seem useful anymore. Broadly speaking, two points of consensus can be found on the issues. First, explicit and implicit knowledge and learning processes exist independently. Most researchers agree that explicit knowledge does not transform into, convert into, or become implicit knowledge (the exception is Rod Ellis, who claims that conversion of explicit knowledge is possible in variational features). Second, explicit knowledge (and learning) is necessary to acquire implicit knowledge and to achieve high ultimate attainment in L2 (except for the earlier claim proposed by Krashen).

One of the most finely articulated differences can be found between N. Ellis and Paradis. N. Ellis claimed that there is a direct effect of explicit knowledge/learning on implicit knowledge, whereas Paradis refuted the idea of a direct effect (interface), but advocated an indirect effect. Paradis (2009) maintained that "explicit instruction has a *direct* effect on explicit language learning and an *indirect* effect on implicit language acquisition… the actual contribution of explicit knowledge to implicit competence is indirect—possibly extensive, but indirect." (p. 69) He thus claimed "there is no interface in any conventional definition of the word." (p. 106) Note that N. Ellis still maintains that there is "an interface, a dynamic one at least." (N. Ellis, personal communication, October 24, 2014). Although they disagree on how

explicit knowledge exerts an influence, they still agree that explicit knowledge and learning are indispensable to the development of implicit knowledge.

As explicit knowledge and learning could potentially play a central role in the development of implicit knowledge, it is useful to understand the development of explicit knowledge through the lens of skill acquisition theory. DeKeyser and others delineated the developmental processes from explicit knowledge to automatized explicit knowledge. It can be reasonably assumed that proceduralized or automatized explicit knowledge can influence the development of implicit knowledge because a more rapid access to explicit knowledge allows L2 learners to process inputs more efficiently. An empirical question can be addressed as to how automatized explicit knowledge influences the acquisition of implicit knowledge of the same structures.

As they are related to the interface issue of explicit and implicit knowledge, the current study also examines the explicit and implicit learning processes. An individual differences approach is taken for revealing the learning process; the study examines the relationship between knowledge and cognitive aptitudes, which offers us a window into observing the underlying learning processes (DeKeyser, 2012). In the next section, literature on cognitive aptitudes for second language learning is reviewed. Further, studies that examined the effects of cognitive aptitudes on L2 attainment will be examined with attention focused on naturalistic SLA settings.

2.8 Cognitive Aptitudes for Second Language Learning

Second language learning aptitudes are conceptualized as a set of cognitive and perceptual abilities that are deployed for various aspects of L2 learning (Carroll, 1981; Linck et al., 2013). Cognitive aptitudes can be regarded as a partly innate trait of mental processing to the extent that it exhibits some degree of stability over long periods of time, and that they predict

future success in L2 learning (Carroll, 1993). For instance, the early aptitude tests such as the Modern Language Aptitude Test or MLAT (Carroll & Sapon, 1959) and Pimsleur Language Aptitude Battery or PLAB (Pimsleur, 1966) were extensively researched during the 1960s and the 1970s and were proven to be reliable and successful in predicting the rate of learning, particularly in intensive foreign language learning contexts (Skehan, 1998).

Since that early research on aptitude tests, significant advancements in cognitive psychology have been made, and a better understanding of language learning and teaching methodologies has influenced reconceptualization of aptitude constructs. Increasing attention has been paid to memory ability, including working memory, as a potential component of foreign language aptitudes (Doughty et al., 2010; Linck, Hughes, et al., 2013; Miyake & Friedman, 1998; Wen & Skehan, 2011). Furthermore, influence from the traditional audio-lingual grammar-based teaching methods in the 1960s and the 1970s narrowed the scope of second language learning aptitude components to explicit types of learning. There is an emerging line of research in educational psychology that focuses on implicit learning and memory processes as new aptitudes for learning (Kaufman et al., 2010; Woltz, 2003). The conceptualization of aptitudes for implicit learning as well as explicit learning appears to be very useful and directly relevant for research that examines two types of L2 learning processes. Specifically, if aptitudes for two types of learning can be stipulated, then the role of these aptitudes can be directly assessed at different stages of L2 development, which allows us to infer the underlying learning processes. Recent research has accumulated evidence for two distinct aptitudes for explicit and implicit learning (Granena, 2013a; Linck, Hughes, et al., 2013), and a few empirical studies have demonstrated that implicit learning aptitudes, measured with the Serial-Reaction Time task, can predict second/foreign language learning outcomes (Granena, 2012, 2013b; Kaufman, et al.,

2010; Linck, Hughes, et al., 2013; Suzuki & DeKeyser, in press). The current research attempts to reveal learning mechanisms that are at work for acquisition of explicit and implicit knowledge by taking an individual differences approach (DeKeyser, 2012; DeKeyser & Koeth, 2010). In the next section, research will be reviewed that has investigated the interactions of individual differences with other variables in the learning process.

2.9 Individual Differences and Ultimate Attainment in Adult SLA

The learning process is unobservable, as opposed to the product of learning, and is often treated as a black box. One approach to exploring learning processes is to infer them from the way in which individual difference variables interact with linguistic and/or contextual variables (DeKeyser, 2012). The assumption is that another variable interacts with an individual difference variable because it requires a mental process that is facilitated or hampered by the individual difference variable. Several studies have documented interactions between starting age of L2 acquisition, linguistic knowledge, aptitudes, and linguistic structures on L2 development, which will be reviewed.

In an early study, Harley and Hart (1997) demonstrated that memory played a more important role for early starters, who received intensive exposure in grade 1, while language-analytic ability predicted the gains in older starters, who received intensive exposure in grade 7, in the immersion program. The results suggest that the different learning processes are determined by age differences, but the caveat was that older learners received a more formal type of instruction, which might have been reflected in the higher predictive validity of analytic ability in older learners.

In their follow-up study, Harley and Hart (2002) investigated older English students (grades 10 and 11) who stayed with a French-speaking family for a three-month exchange

program. This allowed them to observe the learning process in a less formal learning context. The results supported the original immersion program study, showing that analytic language ability was closely related to outcome measures administered at the end of the exchange program, though the relationship was weaker than for the classroom contexts in Harley and Hart (1997).

The studies by Harley and Hart (1997, 2002) focused on the *rate* of learning within a short length of learning periods. Another emerging line of investigations focused on L2 *ultimate attainment* through long-term Length of Residence (LOR) in the target-language country (Krashen, Long, & Scarcella, 1979). The primary aim of these studies was to reveal the learning mechanisms in child SLA and adult SLA.

A study by DeKeyser (2000) investigated acquisition of English syntax (measured with the untimed GJT) by Hungarian immigrants in the United States. The participants had resided in the United States for at least 10 years (average length of residence was 34 years). The study revealed an interesting pattern of correlation between language-analytic ability (measured by the Hungarian version of Words in Sentences in MLAT) and the GJT score. The language score did not correlate with the GJT score for those who arrived before the age of 16, but scores for adult arrivals were statistically significant ($r = .33$, $p < .05$).[13] This indicates the importance of explicit learning for late learners' ultimate attainment, and supports Bley-Vroman's Fundamental Difference Hypothesis (Bley-Vroman, 1990, 2009). According to the hypothesis, child SLA takes place mostly implicitly, through a domain-specific mechanism, whereas adult SLA requires domain-general problem-solving mechanisms for explicit learning, and thus language-analytic ability or explicit learning aptitude plays a role only for adult learners.

[13] Granena (2012) pointed out that the score range was smaller in aptitudes and GJT scores for early learners in DeKeyser (2000).

In order to replicate the interaction between age and aptitudes found in DeKeyser (2000), Abrahamsson and Hyltenstam (2008) examined the hypothesis that only late learners with high grammatical sensitivity will reach near-native levels of L2 proficiency. Highly advanced L2 Swedish with L1 Spanish speakers were recruited and tested on the GJT and the aptitude tests, Swansea LAT.[14] The findings of DeKeyser (2000) were partially replicated: although the correlation for later learners (Age of Onset (AO) > 12) was moderate, it did not reach statistical significance ($r = .53$, $p = .094$).[15] A strong correlation was found between the GJT score and the aptitude tests in early learners whose AO is below 12 ($r = .70$, $p < .001$).[16] Just as DeKeyser (2000) did, however, Abrahamsson and Hyltenstam (2008) found that no adolescent or adults learners scored within the same range as the younger learners unless they scored high on their aptitude measure.

A follow-up study by DeKeyser, Alfi-Shabtay, and Ravid (2010) conducted a cross-linguistic investigation on the role of age and aptitudes; the participants were all Russian L1 speakers, but one group immigrated to the United States acquiring English, and the other group immigrated to Israel acquiring Hebrew. Participants were given a GJT in the L2, which consists of a variety of grammatical structures in Hebrew or English, as well as the same aptitude test in L1 (a Russian version of a verbal academic aptitude test). For the acquisition of English, the correlation for the group with AO <18 was not significant ($r = .11$); that for AO 18–40 was

[14] The composite score of the aptitude tests was used, and the aptitude components tested were phonetic memory, lexical morphological analytical skills, grammatical inferencing skills, aural memory for unfamiliar sound sequences, and the ability to form sound-symbol associations.
[15] This is probably due to "the small number of participants in this group as well as the fact that all 11 late-learners had above-average aptitude" (Abrahamsson & Hyltenstam, 2008, p. 498).
[16] The stimuli sentences used in this Swedish study were extremely difficult (e.g., involving very long and complicated sentences), which might have produced the strong correlation in early learners.

statistically significant (r = .44; p < .05). The same pattern of correlations was found for the acquisition of Hebrew: AO < 18 group (r = -.37, ns) and AO 18–40 group (r = .45, p <.01).

Overall, the studies above provide evidence that explicit learning aptitudes play an important role, especially for adult or late L2 learners. However, recent studies have revealed that explicit aptitudes may not be related to the acquisition of *more automatic* knowledge for late L2 learners (Granena, 2012, 2013b; Granena & Long, 2013). Granena and Long (2013) employed a timed auditory GJT to measure the acquisition of morphosyntax in L2 Spanish; participants were told to make a grammaticality judgment as quickly as possible. They found no significant contribution of aptitude, measured by LLAMA, to the acquisition of morphosyntax in Spanish L2 for late learners (AO 16-29).

In a subsequent study, Granena (2012) attempted to measure linguistic knowledge more rigorously, employing linguistic measures that required automatic to controlled use of language knowledge[17]: the word-monitoring task, the timed auditory GJT, the timed visual GJT, the untimed auditory visual GJT, the untimed visual GJT, and the metalinguistic knowledge test. With these linguistic knowledge measures, Granena (2012) investigated the differential effects of cognitive aptitudes, defined as implicit learning aptitudes (the Serial Reaction Time (SRT) task and LLAMA D) and explicit learning aptitudes (LLAMA B, E, F, and the intelligence test GAMA) on the acquisition of L2 Spanish morphosyntactic knowledge. Consistent with the findings of Granena and Long (2013), automatic use of language was *not* related to explicit aptitudes for late learners (AO>16), while automatic use was significantly predicted by the

[17] From the present dissertation's point of view, one of her tasks (i.e., the word-monitoring task) can be considered as an implicit knowledge measure, but we present her tasks following her operationalizations.

implicit learning aptitudes.[18] Granena (2012) further revealed that the controlled use of language was related to explicit aptitudes.[19]

The fact that explicit learning aptitude was not related to L2 knowledge seems to contradict the findings of the studies by DeKeyser and Abrahamsson and Hyltenstam. This is probably, as Granena and Long suggested, due to the differences in the linguistic knowledge measures.[20] Granena's studies employed the tasks that tapped into automatic use of L2 knowledge (i.e., timed GJT and the word-monitoring task), while the DeKeyser and Abrahamsson and Hyltenstam's studies used the *untimed* GJTs that might draw more explicit knowledge. These findings suggest that the contributions of aptitudes vary depending on the types of linguistic knowledge that are measured. With more extensive test battery of explicit and implicit knowledge, the present study investigates the extent to which explicit and implicit learning aptitudes contribute to the acquisition of explicit and implicit knowledge.

[18] An unexpected finding was that automatic language use (i.e., the word-monitoring task) was related to explicit aptitudes only in early learners (AO 3-6).

[19] It is noted that the large-scale study of Granena (2012) had multiple measures and not all the correlations were consistent with the interpretations above. For instance, the correlations between the timed auditory GJT and the explicit aptitudes were .27 ($p = .056$) in early learners and .26 ($p = .075$) in late learners.

[20] In addition to the differences in the knowledge measures, one could also attribute the conflicting findings to different measures of aptitudes used across the studies. However, aptitudes tests are probably not the source of the difference, because different aptitude tests are used, such as the MLAT in DeKeyser (2000), an equivalent of the verbal SAT in DeKeyser et al. (2010), and the LAT aptitude test in Abrahamsson and Hyltenstam (2008). LAT is a former version of the LLAMA tests, which is used in Granena's studies. LLAMA has been validated as a measure of explicit learning aptitudes (parts B, E, and F) and implicit learning aptitudes (part D) with satisfactory test-retest reliability (Granena, 2013a).

Chapter 3: Motivations for the Current Study

The present dissertation research was motivated by the two related gaps in the body of literature on the issue of explicit and implicit knowledge and learning: (1) lack of valid measurements for implicit knowledge and (2) dearth of empirical investigations into the interface issue. First, existing measures for implicit knowledge appeared to be too coarse to assess implicit knowledge; they might be considered to be measures for automatized explicit knowledge, not implicit knowledge (Suzuki & DeKeyser, in press). The present study addressed this gap by devising more fine-grained tasks that can assess implicit knowledge more validly, and compared them with the existing measures for implicit knowledge. In other words, the first goal of this dissertation was to validate the behavioral measures that could tap into implicit knowledge and automatized explicit knowledge separately. The measurements used here for implicit knowledge assessment were inspired by the psycholinguistic research in which real-time predictive sentence processing have been examined. Specifically, the eye-tracking-while-listening paradigm (i.e., visual-world paradigm) was employed to measure implicit knowledge as well as other RT-based methods, i.e., a word-monitoring task and a self-paced reading task. This set of measurements was compared against the three measures that impose time-pressure (i.e., the time-pressured auditory GJT, the time-pressured visual GJT, and the time-pressured fill-in-the-blank test).

In order to validate the measurements for implicit knowledge and automatized explicit knowledge, construct validity was assessed more rigorously via three statistical procedures: Confirmatory Factor Analysis (CFA), Multi-Trait Multi-Method (MTMM) analysis, and Structural Equation Modeling (SEM). These analyses were statistically powerful enough to estimate the stability of latent constructs by taking into account the measurement errors, rather than zero-order correlations between measurements.

The CFA analyses primarily focus on the validation of the two-factor model in Figure 3, which contrasts with the one-factor model in Figure 4. It was hypothesized that the two-factor model would fit better than the one-factor model. More specifically, we investigated to what extent the six measurements exhibited evidence for convergent and discriminant validity. Convergent validity refers to the extent to which different measures of the same trait tend to cluster together, whereas the discriminant validity referred to the extent to which measures of different traits, using either the same or different test methods, tend to diverge (Bachman, 1990). In the present study, convergence concerns the extent in which the measurements utilizing the online psycholinguistic technique, which could investigate real-time sentence processing, tapped into the hypothesized implicit knowledge construct. It also referred to the extent in which the measurements utilizing the time-pressured form-focused techniques, requiring focus on form, tapped into a distinct automatized explicit knowledge construct. Discriminant validity, in the current study, referred to a zero or weak relationship between the set of measurements for implicit knowledge and those for automatized explicit knowledge.

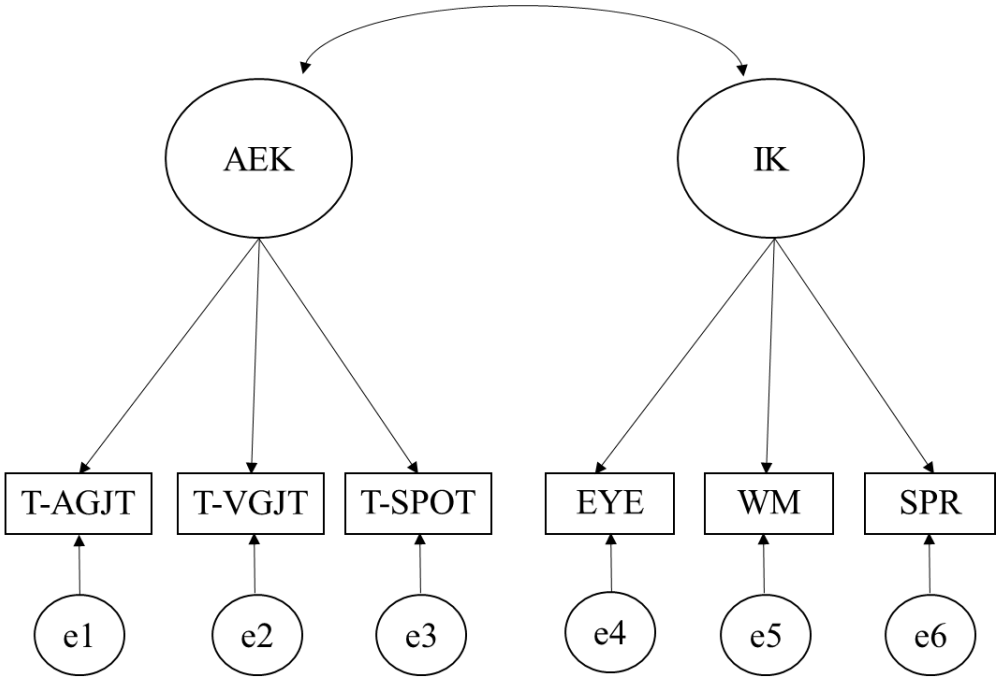


Figure 3. CFA Model 1: Two-factor Model

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT[21], Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task.

[21] SPOT is a Simple Performance-Oriented Test in which participants fill in the blank with a target grammatical structure in a written sentence (see Methods).

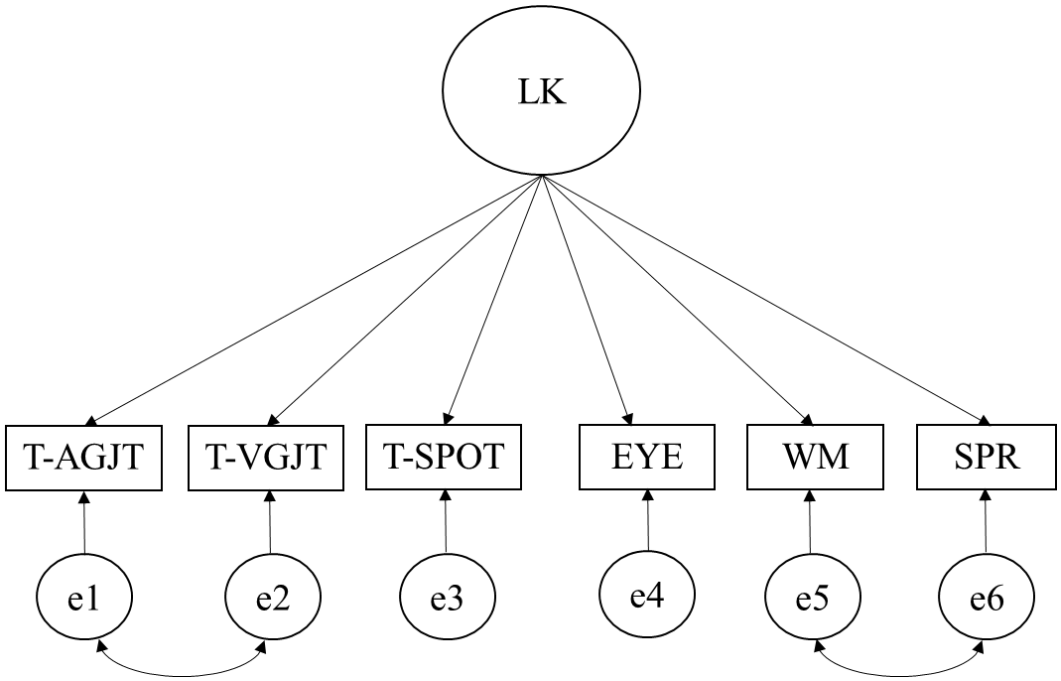


Figure 4. CFA Model 2: One-factor Model

Note. LK = Linguistic Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

In order to investigate the construct validity of measurements more rigorously, a CFA model of the multitrait-multimethod (MTMM) analysis was also conducted to examine patterns of both convergence and discrimination among correlations of the measures. The theoretical framework for MTMM analysis was first proposed by Campbell and Fiske (1959); the statistical procedure utilizing CFA was formalized by Widaman (1985). The key advantage of MTMM analysis is that it assesses the extent to which the traits were measured validly by taking into account the method artifacts. This approach has been utilized in the second language assessment field (Bachman & Palmer, 1982; Buck, 1992; Llosa, 2007; Sawaki, 2007) to assess the extent to which variance in the measurements could be attributed to latent constructs of linguistic

knowledge (traits) and to specific methods (methods effect). The current study utilized a pair of measurements that shared very similar methods; the word-monitoring task and the self-paced reading task used the reaction time during the real-time sentence processing for comprehension, whereas the visual GJT and the auditory GJT shared the same procedure except for the modality difference. The question pertaining to construct validity was whether the traits (explicit and implicit knowledge) could be measured rather than the method effects.

After the two hypothesized constructs were modeled both in CFA and MTMM analyses, the nature of the two constructs was scrutinized by SEM analysis. As the good fit of a two-factor model to the data cannot provide conclusive evidence that the two traits are explicit and implicit, SEM analysis was conducted to address this issue by adding the two cognitive aptitudes, for explicit and implicit learning, as predictors of the two latent factors in the two-factor model. If the hypothesized implicit knowledge factor was implicit knowledge, it should be predicted by the implicit learning aptitude more strongly than by the explicit learning aptitude (See Suzuki, 2013 for a similar approach). In contrast, explicit learning aptitude should predict the hypothesized automatized explicit knowledge factor more strongly than implicit learning aptitude does.

Second, it is conceivable that the scarcity of empirical research on the interface issue of explicit and implicit knowledge and learning stems from the lack of valid measurements for explicit and implicit knowledge (Dörnyei, 2009; Hulstijn, 2002). As shown above, R. Ellis's seminal work is the starting point of our endeavor to investigate the issue, but further refinement of measurement seems to be needed (Suzuki & DeKeyser, in press). Only with clearly valid measurements of explicit and implicit knowledge can we proceed to tackle this fundamental issue in the SLA field. After providing evidence for the validity of implicit and explicit measures in the current study, we proceeded to examine the interface issue empirically.

Figure 5 lays out the L2 learning processes pertaining to the interface issues. Most SLA researchers agree that there are distinct constructs of explicit and implicit knowledge, and that explicit knowledge exerts some influence on the acquisition of implicit knowledge. It is, however, an open question exactly how explicit learning and knowledge influence the acquisition of implicit knowledge. Using SEM analyses, two related questions were addressed separately at the learning and knowledge level.

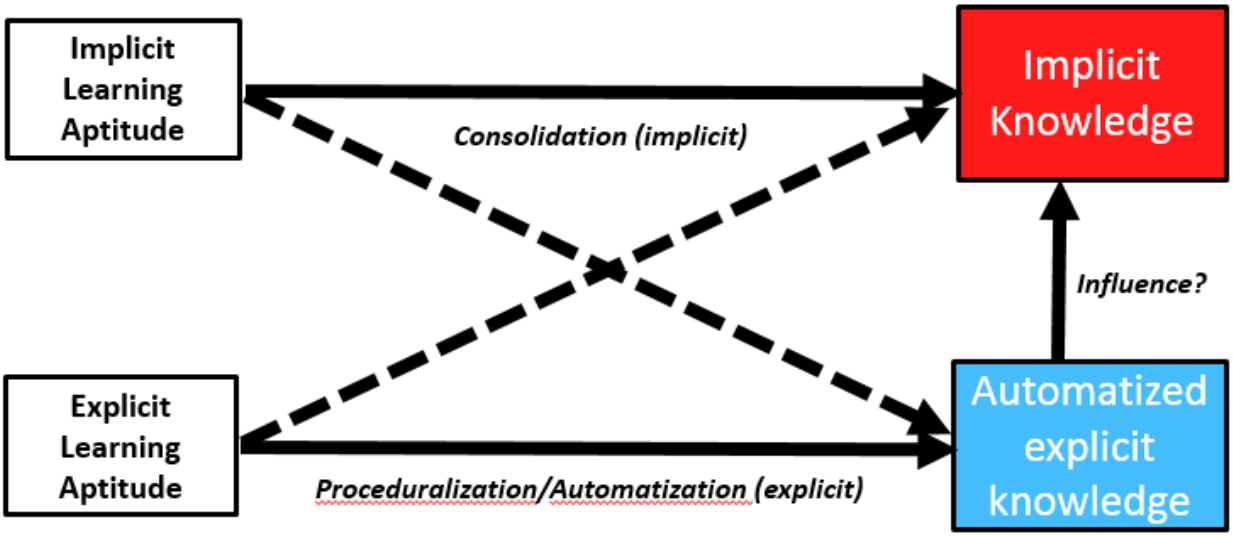


Figure 5. Summary of implicit/explicit knowledge interface issues

First, the current study explored the role of learning processes by examining the role of cognitive aptitudes for explicit and implicit learning on the acquisition of two types of knowledge. Previous research has investigated the role of cognitive aptitudes for explicit and implicit learning in immersion L2 environments; it is unclear, however, as to which types of linguistic knowledge were correlated with individual differences in aptitudes, due to the ambiguous nature of the language tests. Using SEM analyses, the current study investigates the contributions of the aptitudes to automatized explicit and implicit knowledge. It is assumed that individual differences in a cognitive aptitude for one type of learning processes (e.g., explicit

learning) are more closely related to the product of that learning (e.g., explicit knowledge). It is straightforward to assume that an implicit learning aptitude is related to implicit knowledge and that an explicit learning aptitude is related to automatized explicit knowledge. However, it is a more open question of how one type of aptitude (e.g., explicit) was involved in the acquisition of the other type of knowledge (e.g., implicit). In the real world, the state of automatized explicit knowledge and implicit knowledge vary greatly among L2 learners, and the contribution of aptitudes may also change depending on the stage of learning. For instance, the initial stage of explicit learning (proceduralization) and the later stage (automatization) may draw on different kinds of aptitudes. Note that the interpretations of the relationship between aptitude and knowledge are not conclusive but rather are considered to be supplementary to the understanding in the relationship between the two types of knowledge.

Second, the direct relationship between the products of the two learning processes, automatized explicit and implicit knowledge, was investigated. There is no empirical research on whether automatized explicit knowledge (the product of explicit learning) exerts influence on the acquisition of implicit knowledge. The current study aimed at shedding light on the potential link between automatized explicit knowledge and implicit knowledge.

Another component of aptitude, phonological short-term memory, was examined as a predictor of the acquisition of explicit and implicit knowledge, because it is one of the potential mediating factors for explicit and implicit learning. Since it can be assumed that higher memory ability opens a larger window to process language sequences for inducing grammatical rules (at least in adults), phonological short-term memory is expected to correlate with both aptitudes for explicit and implicit learning (see Janacsek & Nemeth, 2013 for a more detailed review on the relationship between memory and sequence learning).

Little research has investigated the role of phonological short-term memory on the development of explicit and implicit knowledge respectively (see Linck, Osthus, Koeth, & Bunting, 2013 for review on the role of phonological short-term memory and working memory in SLA). Phonological short-term memory had been found to play a significant role in L2 grammar learning in laboratory-based and classroom-based research (e.g., French& O'Brien, 2008; Martin & Ellis, 2012; Williams, 1999; Williams & Lovatt, 2003). Previous studies utilized outcome measures that are more likely t athercole & Baddeley,1993o allow for accessing explicit knowledge, and we speculated that phonological short-term memory plays a role particularly in the acquisition of automatized explicit knowledge. On the other hand, it is much less unclear whether it also plays a role in the acquisition of implicit knowledge; the question is left open as to whether phonological short-term memory covaries with the acquisition of implicit knowledge. In sum, the present study investigates to what extent phonological-short-term memory, in combination with the explicit learning aptitude and the implicit learning aptitude, contributes to the acquisition of morphosyntax.

Chapter 4: Research Questions and Hypotheses

The first two research questions investigated to what extent the new test battery could tap into the two constructs of automatized explicit knowledge and implicit knowledge separately. Seven hypotheses were put forth to examine the research questions in detail.

1) Does the test battery tap into the constructs of implicit knowledge and automatized explicit knowledge distinctly? Specifically, do the three online measures (i.e., visual-world task, word-monitoring task, and self-paced reading task) tap into the construct of implicit knowledge and do the existing measures for "implicit knowledge" (i.e., timed form-focused tests) tap into automatized explicit knowledge?
    - Hypothesis 1: The data structure of the six measurements demonstrates a good fit to the two-factor model.
    - Hypothesis 2: The factor loadings are strong and significant (systematic) for automatized explicit knowledge and implicit knowledge (convergent validity).
    - Hypothesis 3a: The relationship between the two latent factors is insubstantial (discriminant validity).
    - Hypothesis 3b: The data structure of the six measurements demonstrates a poor fit to the one-factor model.
    - Hypothesis 4: The covariance between the similar measurement methods is smaller than the covariance between the measurements for the traits (method effects).
    - Hypothesis 5: The latent factor hypothesized as implicit knowledge is predicted by the cognitive aptitude for implicit learning more strongly than the cognitive aptitude for explicit learning.

- Hypothesis 6: The latent factor hypothesized as automatized explicit knowledge is predicted by the cognitive aptitude for explicit learning more strongly than the cognitive aptitude for implicit learning.

The second research question and the corresponding hypothesis 7 were driven by the theoretical motivation and empirical findings. In theory, "implicit memories are acquired slowly because this requires a very large number of encounters with each particular form" (Paradis, 2009, p. 95). This was empirically confirmed by Suzuki and DeKeyser (in press), who found a positive relationship between the aptitude for implicit learning and L2 knowledge only when L2 speakers had more L2 experience in a naturalistic acquisition setting.

2) Does the amount of L2 exposure in the immersion setting, estimated by the length of residence, change the results of hypotheses 1 to 6?
    - Hypothesis 7: The results from L2 speakers who received long-term exposure in an immersion setting confirm hypotheses 1-6 more convincingly than the results from L2 speakers with less naturalistic exposure

The following four (3, 4, 5, and 6) questions investigated the interface issues of explicit and implicit knowledge and learning. The third research question focused on the relationship between explicit and implicit *knowledge* and investigated the extent to which the automatized explicit knowledge influenced the acquisition of implicit knowledge.

3) To what extent does the automatized explicit knowledge influence the acquisition of implicit knowledge?
    - Hypothesis 10: Automatized explicit knowledge contributes to the development of implicit knowledge.

The fourth and fifth research questions focused on the contributions of explicit and implicit *learning* to the development of explicit and implicit knowledge. These two questions were addressed by examining the effects of cognitive aptitude for explicit and implicit learning on the acquisition of automatized explicit and implicit knowledge.

4) To what extent does explicit learning aptitude contribute to the acquisition of automatized explicit and implicit knowledge?
    - Hypothesis 8a: Explicit learning aptitude plays a role in the acquisition of automatized explicit knowledge (based on skill acquisition theory).
    - Hypothesis 8b: Explicit learning aptitude plays a role in the acquisition of implicit knowledge.
5) To what extent does implicit learning aptitude contribute to the acquisition of automatized explicit and implicit knowledge?
    - Hypothesis 9a: Implicit learning aptitude does NOT play a role in the acquisition of automatized explicit knowledge.
    - Hypothesis 9b: Implicit learning aptitude plays a role in the acquisition of implicit knowledge (Granena, 2013b; Suzuki & DeKeyser, in press).

The last research question addressed whether phonological short-term memory could moderate the effects of explicit and implicit learning.

6) To what extent does phonological short-term memory moderate the effects of explicit and implicit learning aptitudes on the acquisition of explicit and implicit knowledge?
    - Hypothesis 11a: Phonological short-term memory plays a role in the acquisition of automatized explicit knowledge.

No prediction was made for the role of phonological short-term memory in the acquisition of implicit knowledge.

Chapter 5: Methods

5.1 Participants

One hundred Japanese second language speakers (29 male and 71 female), whose first language was Chinese, were recruited in Tokyo and the surrounding Kanto area. Fifty-one NSs were also recruited to serve as a baseline for the linguistic knowledge tasks.[22] Four requirements had to be met by L2 speakers in order to participate in the study: (1) proficiency, (2) age of arrival in Japan, (3) length of residence, and (4) educational background. First, only advanced-level Japanese L2 speakers were recruited. They were screened for Japanese proficiency, which had to be equivalent to or higher than N1 in the standardized Japanese Language Proficiency Test (JLPT).[23] Second, we only focused on late L2 speakers, who arrived in Japan after the age of 17. Third, their LOR in Japan was two years or longer. This cut-off point for LOR was roughly based on the previous findings that implicit knowledge seems to be exhibited most efficiently in online measurements (i.e. the word-monitoring task) when L2 speakers have been immersed in the target country for two and half years of residence or longer (Suzuki & DeKeyser, in press). Fourth, participants possessed at least a bachelor's degree or were currently enrolled in a four-year college.

Detailed background information about the L2 speakers is presented in Table 4. The sampled population consisted of undergraduate students (n = 34), MA students (n = 40), PhD students (n = 14), and office workers (n = 12). During their undergraduate studies, 43

[22] An initial group of thirty-one participants took the entire linguistic test battery. After the initial testing, we found that the visual-world task did not work as expected. Another group of 20 participants was tested only with the revised visual-world task.

[23] "JLPT N1 (which corresponds to the previous JLPT Levels 1) is roughly equivalent to the ACTFL Superior on the OPI scale (Kanno, Hasegawa, Ikeda, & Ito, 2005). JLPT Level 1 is the minimum requirement for acceptance into a regular college undergraduate/graduate program in Japan

participants majored in a Japanese-related field (i.e., Japanese, Japanese education, linguistics), whereas 57 participants majored in other fields of study (n = 57).

Table 4. *Background Information of the L2 Speakers*

	Age at Testing	Age of Arrival	Length of Residence (months)	Onset of Instruction	Length of Instruction (months)
Mean	25.97	21.36	47.29	19.01	41.11
SD	4.47	2.66	27.71	2.25	17.44
Min	19	17	24	13	3
Max	47	30	197	27	84

Note: None of the variables were normally distributed according to the Kolmogorov-Smirnov test ($p < .05$)

5.2 Target Structures

Four Japanese linguistic structures were tested across the six linguistic knowledge measurements: 1) transitive/intransitive verb pairs, 2) classifiers, 3) locative particles (*ni/de*), and 4) the *tameni/youni* construction indicating purpose. We chose these four structures because they generate some prediction of upcoming information, which can be demonstrated by the visual-world task.

5.2.1 Transitive/Intransitive Verb Pairs. Japanese has approximately 350 sets of transitive-intransitive verb pairs (Jacobsen, 1992). The pairs are similar in form; the transitive verb, *yaburu* (to tear), has an intransitive counter-part, *yabureru* (to be torn), for instance. Some generalizations can be made regarding the morphological features: (1) the verb ending with *aru* is intransitive, and it can be converted to the transitive verb changing *aru* to *eru*; (2) the verb ending with *reru* is always intransitive; and (3) the verb ending with *–su* is always transitive (Iori, Takanashi, Nakanishi, & Yamada, 2000). The productivity of the rules, however, is limited because for most verbs the classes cannot be determined solely based on the morphological form, and some classes have only a few verbs that form a class, i.e., low-type frequency. This implies that the transitivity of verbs must be learned item by item. As would be expected given these

characteristics, it has been found in L2 research that these verb pairs are hard to acquire (Lin, 2002) and that they are learned by item-based learning mechanisms (Nakaishi, 2005).

As shown in sample sentence 1, a theme is discernible by the object-marking particle *o* for the transitive verb (i.e., *ageru*; to raise). If the subject-marking particle *ga* isz used, the sentence becomes ungrammatical for the transitive verb. Sentence 2 demonstrates the intransitive verb usage. The subject should be marked with the subject-marking particle *ga* rather than *o*. Sixteen transitive/intransitive verb pairs were chosen for the current study; the average frequency of the transitive verbs was 16.82 per million, and that of the intransitive verbs was 17.70 per million (see Appendix A). No significant difference exists between them according to a t-test ($p > .1$).

1. *Kyuryou o/*ga ageru to, hataraku hito wa ganbaru.*
   Salary-OBJ raise if, work person-TOPIC work harder.
   (If you raise the salary, workers will work harder.)
2. *Ookii jiken ga/*o okiru to, kanarazu shinbun ni deru.*
   Big case-SUB happens when, always newspaper-location appears
   (When a big case happens, it always appears in the newspaper.)

5.2.2 Classifiers. Numerical classifiers are used when counting objects with clear delineating boundaries (Iwasaki, 2002). In Japanese, there are over 150 classifiers, but approximately 30 of them are regularly used. There are more varieties of classifiers in Chinese, but some classifiers are shared across the two languages (e.g., the same character with a similar pronunciation, 頭 (*tou*) and 头(*tóu*), is used for counting horses in Japanese and Chinese, respectively). Classifier-noun pairs, which are not shared between the languages, were chosen for the study. Eight classifiers are chosen: the first four classifiers in Japanese in Table 4 are not used as a classifier in Chinese, whereas the other four classifiers do exist in Chinese, but the

nouns matched with them are different across the languages. For instance, 台 (dai/tái) is a classifier in both languages, but Japanese nouns chosen in the study are not used with 台 in Chinese. There were 32 classifier-noun pairs chosen for the study; each classifier was matched with four different nouns (Appendix B).

Table 4

Usage of Classifiers

Classifier	Usage in Japanese	Chinese equivalent
着 (*chaku*)	Cloth	No
軒 (*ken*)	House, store, etc.	No
足(*soku*)	Shoes, socks, etc.	No
羽 (*wa*)	Birds	No
台 (*dai*)	Vehicle, machines, etc.	Yes
枚 (*mai*)	Thin flat objects	Yes
冊 (*satsu*)	Books	Yes
匹 (*hiki*)	Small animals, fish, etc.	Yes

5.2.3 Locative Particles: *Ni*/*De*. The particles *ni* and *de* are multifunctional case markers, and the usage for locations was focused on in the current study. In particular, *de* indicates the place where an action takes place, whereas *ni* is used to indicate the place where a thing or a person exists. In other words, the particle and the state of the verb need to agree with one another. Since Chinese does not have this distinction, the usage of *ni*/de is often problematic especially for L2 Japanese learners with L1 Chinese. It has been found that Chinese speakers tend to overuse *ni* for *de* (Hasuike, 2004, 2012). The sample sentences 3 and 4 illustrate the differences between the two particles. Not all of the usage for *ni* is difficult, and a relatively easier usage is expressing destination with motion verbs (e.g., *cafe ni hairu*; enter the cafe). This usage of *ni* is explained in more detail for the visual-world paradigm, which will be explained

below. In sum, action verbs agree with the location particle *de*, static verbs with the location particle *ni*, and motion verbs with the destination particle *ni*.

3. *Koohi wa makudonarudo de nomu koto ga ooi*.
   Coffee-TOPIC McDonald-LOCATION drink is often the case
   (It is often the case that I drink coffee at McDonald's.)
4. *Takai apaato wa eki no mae ni aru.*
   Expensive apartments-TOPIC station GENITIVE front-ni located.
   (Expensive apartments are located in front of the station.)

5.2.4 Conjunctions Indicating Purpose: *Tameni*/*Youni*. In Japanese, purpose is expressed by an adverbial clause ending in *tameni* or *youni*. The usage of them is constrained by the state of the predicate in the adverbial phrase and by whether different subjects are allowed in the adverbial and main clauses (Table 5). *Tameni* is used to indicate purpose of actions or events that are volitional, whereas *youni* expresses non-volitional purposes. In addition, *youni* permits different subjects in the subordinate and the main clause, whereas *tameni* does not. In other words, if the event is not controlled by the subject of the main clause, then *youni* is preferred (Maeda, 2006). The current study focuses on the distinction between *tameni* and *youni* in the adverbial clause containing volitional verbs; that is, whether the subject in the main clause can be different from the one in the adverbial clause (sample sentence 5). It has been reported that this distinction is hard to acquire by advanced L2 Japanese speakers because purpose is expressed only by 為了(*wèile*) in Chinese (Inagaki, 2009).

Table 5. *Usage of Tameni and Youni*

		Subjects in main/adverbial clause	
		Same	Different
predicate in adverbial clause	volitional	*Tameni*	*youni*
	non-volitional	*Youni*	*youni*

5. *Shiken ni musuko-ga goukaku suru youni(*tameni), kazoku ga ouen shita.*
   exame-ACC son-SUB pass purpose family-SUB cheered
   (So that the son passes the exam, the family cheered him.)

5.3 Instruments

A battery of nine measures was administered to participants. Six of the tests were linguistic knowledge measures regarding the four linguistic structures described above. The three online-sentence processing measures (the visual-world task, the word-monitoring task, and the self-paced reading task) were hypothesized to tap into implicit knowledge, whereas the other form-focused measures (the timed auditory GJT, the timed visual GJT, and the timed fill-in-the-blank test) were hypothesized to draw on automatized explicit knowledge. As shown in Table 6, the crucial differences between the two types of measures lie in 1) real-time anticipatory sentence processing and 2) focus of attention. All three online measurements examine whether test-takers can incrementally process the sentence while their attention is focused on the meaning of the sentences. In contrast, the two types of GJTs and the fill-in-the-blank test require them to focus on form or grammatical target points under time pressure.

Three additional cognitive aptitude tests were administered. The first test, LLAMA F, measures linguistic inductive ability, which was operationalized as explicit learning aptitude. An aptitude for implicit learning was operationalized as sequence pattern learning or inductive learning without awareness, which was measured with the Serial-Reaction Time Task. In addition to the cognitive aptitude measures for explicit and implicit learning, a measure for phonological short-term memory (letter span task) was administered to examine the role of memory on L2 acquisition.

Table 6. *Task Features of the Linguistic Knowledge Measurements*

	Visual World	Word Monit.	Self-paced	Timed AGJT	Timed VGJT	Timed fill-in-the-blank (SPOT)
Data	Eye	RT	RT	Accuracy	Accuracy	Accuracy

	Fixation Proportion					
Real-Time Processing	Yes	Yes	Yes	No	No	No
Focus	Meaning	Meaning[a]	Meaning	Form	Form	Form
Time Pressure	No	Yes	Yes	Yes	Yes	Yes
Modality	Aural	Aural	Written	Aural	Written	Written

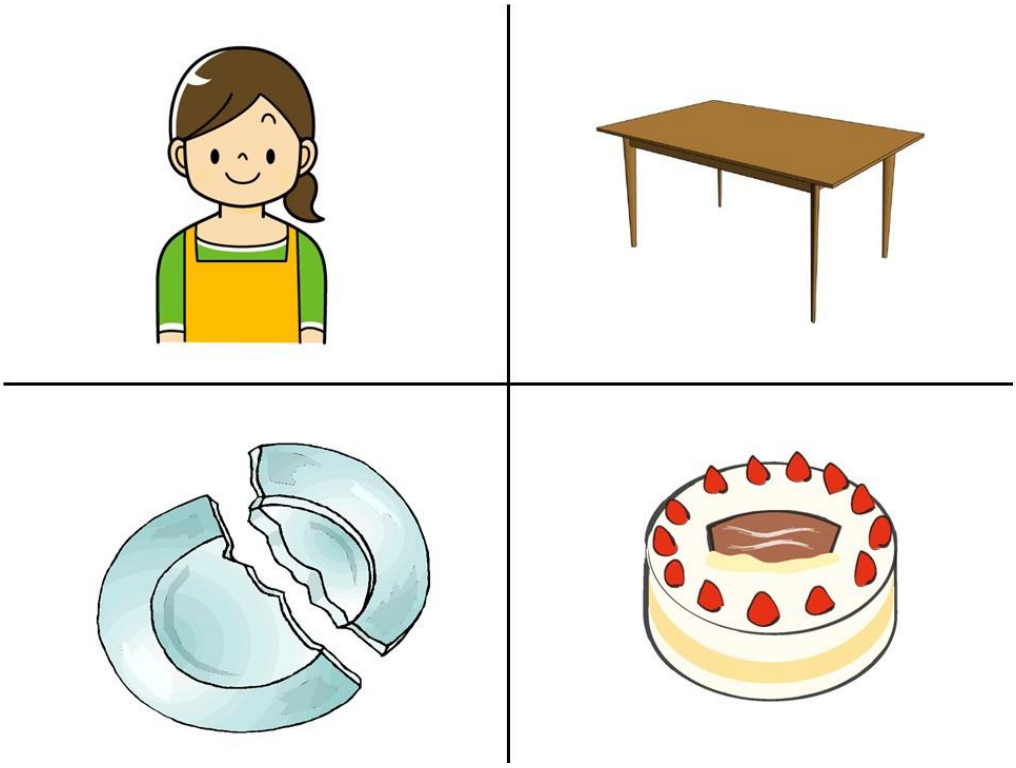
a. The focus of attention is also directed to the monitoring word.

5.3.1 Visual World Paradigm. In the visual-world task, participants were first presented with a scene consisting of four pictures on the computer screen for 5.5 seconds. They then listened to sentences while their eye-movements were being tracked. Sixteen visual scenes or trials were prepared for each of the four linguistic structures tested, and participants heard two sentences for each trial (2 sentences*16 trials*4 structures = 128 sentences). The critical sentence was always presented as the first sentence so that participants were not influenced by any information from the previous sentence. The second sentences acted as fillers to divert the participants' attention from the critical sentence and to avoid revealing the purpose of the study. After each trial, a yes/no comprehension question was asked to ensure that participants' attention was focused on the meaning of the sentence (cf. Dussias et al., 2013). Half of the questions asked about the critical sentence, and the other half about the filler sentence. There were two types of trials for each target structure (i.e., eight trials for each trial type), which will be delineated below. Eye movements were recorded using an EyeLink-II system (SR Research, Osgoode, Ontario, Canada) with a sampling rate of 500 Hz. Two practice trials were given to familiarize the participant with the procedure of the task. Note that the location of the four objects on display was rotated across trials.

5.3.1.1 Transitive/Intransitive. In order to test transitive/intransitive structures, a visual scene was constructed consisting of a theme (e.g., the broken dish), a person (e.g., the mother), a contrast object (e.g., the table), and a distractor (see Figure 6). Two types of critical sentences

were aurally presented: transitive and intransitive sentences. The first part of both sentences always followed the same form: *NP1-ACC-transitive verb-iru-no-wa-adverb-NP2* (It is NP2 that TRANSITIVE-VERB NP1) and *NP1-SUB-intransitive verb-iru-no-wa-adverb-NP2* (The reason is NP2 why NP1 INTRANSITIVE-VERB) (Figure 6). In the transitive trials, NP2 was always a person (e.g., the mother), whereas in intransitive trials it was always a contrast object (e.g., the table). This manipulation was important because if only a person is mentioned in the stimulus sentences, participants might start to rely on a strategy where they look at a person irrespective of linguistic input.

The region of interest was from the onset of the case markers (*ga* or *o*) to the onset of NP2. Note that the adverb was inserted between the particle *wa* and NP2 in order to create a longer region of interest. If participants were sensitive to the transitivity of the verb, then the looks to the person (e.g., mother) would be greater in the transitive trials than in the intransitive trials. This is because a segment of NP-ACC and *te*-form of a transitive verb (i.e., *osara-wo watte*) implied an action doer.



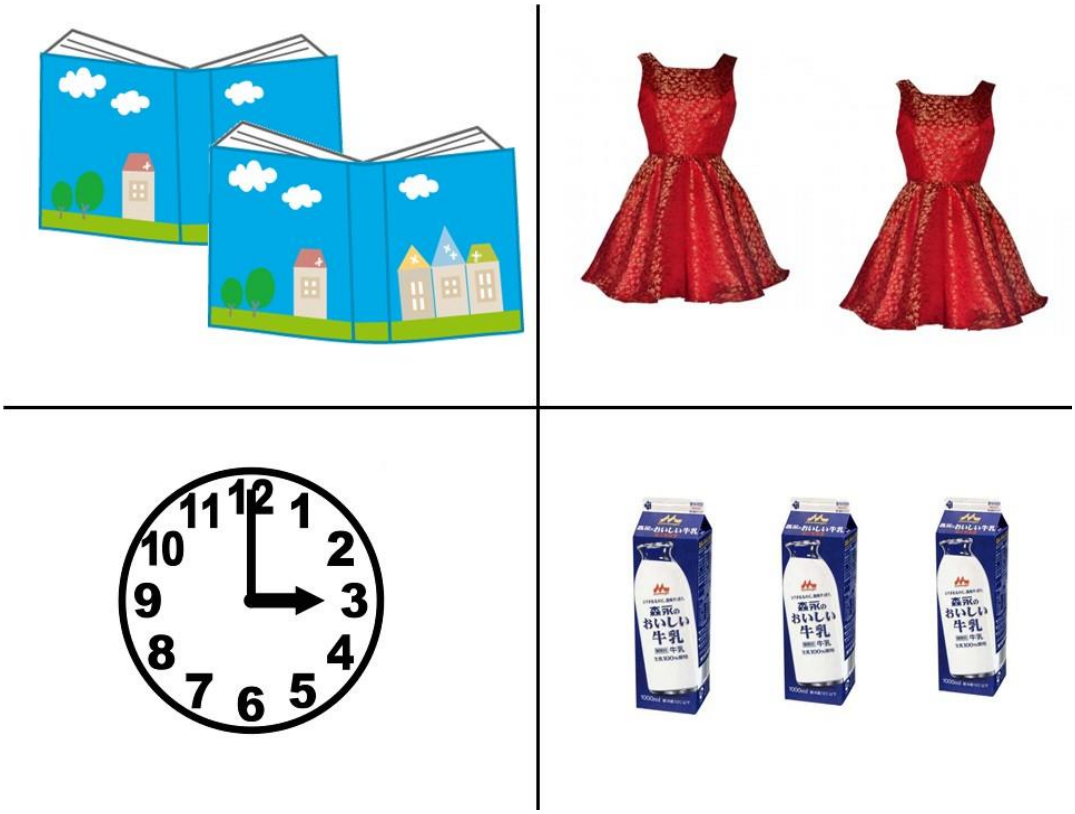
Transitive: *Osara* ***wo watte iru no*** *wa soko ni iru okaasan desu.*
Dish-ACC breaking be NOMINALIZER TOP there-LOC exist mother be.
(It is the mother that is breaking the dish.)
Intransitive: *Osara* ***ga warete iru no*** *wa* ***soko ni*** *aru teeburu kara ochite shimatta kara desu.*
Dish-SUB breaking be NOMINALIZER TOP there-LOC exist table from fall off because.
(The dish is broken because it fell off the table.)
*Region of interest is bolded and underlined.

Figure 6. Visual Scene and Critical Sentences for Transitive/Intransitive Structure

5.3.1.2 Classifiers. A visual scene for the classifier condition consisted of four objects: a target noun (e.g., two dresses), a competitor noun (e.g., two picture books) and two other distractors (see Figure 7). The target noun was defined as the object for which we measured the percentage of fixations as dependent variables (e.g., dresses). The number of the target and competitor nouns was always equal (varying from 1 to 4), and the number of distractor objects was always different from them.

The critical sentence always formed the following order: *Number-Classifier-Genitive-[relative clause]-Noun*. The participants heard two types of critical sentences across trials. In the classifier-matched trials, the classifier (e.g., *chaku*) is matched to the target word (e.g., dress), the classifier (e.g., *satsu*) is matched to the competitor noun in the classifier-mismatched trials. Two counter-balanced lists were created, such that one target picture was referred to in a list, whereas the other target noun in the same display was referred to in the other list. In other words, a target picture always became a target and a competitor once in the same display across the two lists.

The region of interest was from the onset of a classifier to the offset of the relative clause. The relative clause was inserted to create a longer buffer region before participants could predict the forthcoming noun before they heard it. If participants could use the information of the classifier incrementally, then the looks to the target noun (e.g., dresses) would be greater when hearing the matched classifier than the mismatched classifier.



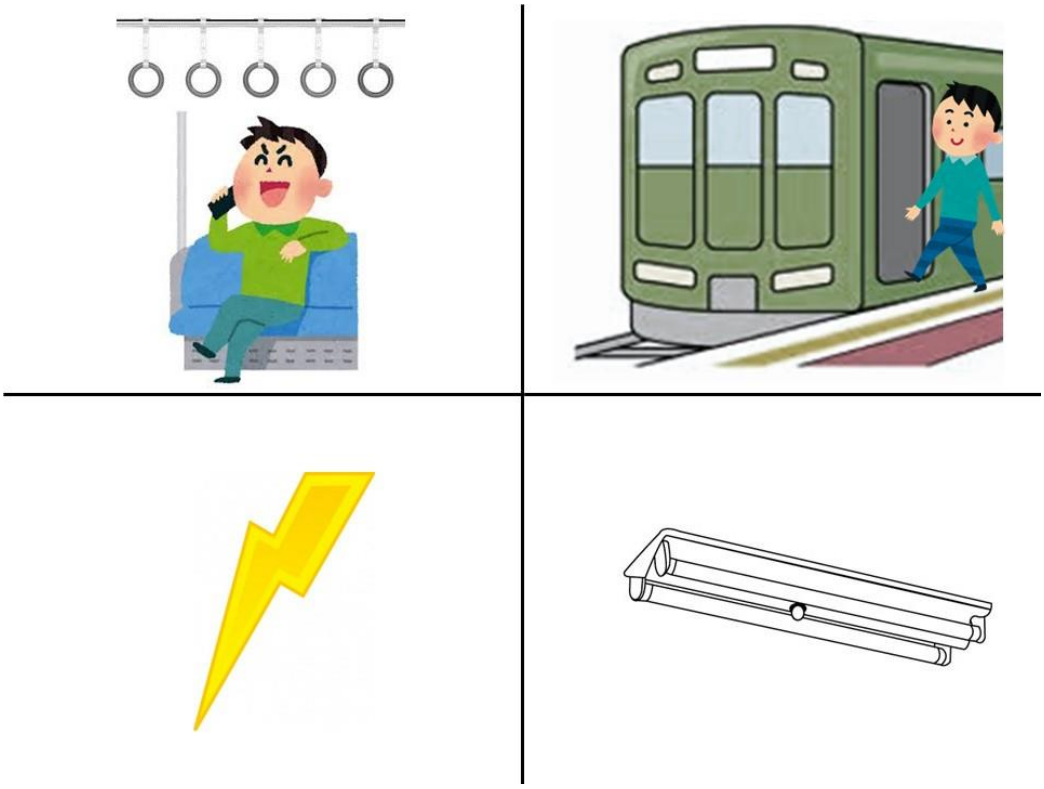
Classifier-matched Trials (when the target noun is dress)
*Ni-**chaku no naraberareta** doresu ga kono heya ni arimasu.*
Two-CHAKU GEN laid out dress-SUB this room-LOC exist.
(There are two dresses in this room)
Classifier-mismatched Trials (when the target noun is dress)
*Ni-**satsu no naraberareta** ehon ga kono heya ni arimasu.*
Two-SATSU GEN laid out picture book-SUB this room-LOC exist.
(There are two picture books in this room)

Figure 7. Visual Scene and Critical Sentences for Classifier

5.3.1.3 Ni/De. A visual scene for the locative particles, *ni/de*, consisted of person A (performing an action in a location), person B (getting to the same location as Person A), and two distractor objects (Figure 8). The location depicted in the pictures was the same between person A and person B (e.g., train). The destination particle *ni* links motion verbs toward the location (e.g., get on, run to, enter), whereas the location particle *de* links the action verbs at the location (e.g., talk, clean). Instead of using the *ni* for the location of existence, we used the usage

of destination because motion verbs were expressed and interpreted more easily in the picture. In the pilot study, pictures with static objects/person were used to denote the location for existence in contrast to pictures with action doers; however, the native speakers did not show the distinction between them in their eye-movement data. It was thus decided that the destination usage of *ni* was used to compare with *de*, in order to infer whether listeners could distinguish the locative usage of *ni* and *de*. The pictures were shown to six Japanese NSs, who did not participate in the actual experiment, and they were asked to describe the pictures in Japanese. Sixteen pairs of pictures were chosen that were described as action location and destination by 5 of the 6 people.

The first part of the sentence always formed the following pattern: *NP-de/ni-Adverb-VP*. The region of interest was from the onset of the particle, *ni* or *de*, to the onset of the VP. The adverbs were included to create a buffer region for examining whether participants could use the particle's information before they heard the verb. If participants were sensitive to the locative particle, then the looks to Person A would be greater when hearing the *de* than the *ni*. If participants were not sensitive to the locative particle (i.e., L2 speakers who overuse *ni* for *de*, see Hasuike, 2004. 2012), the proportion of looks to Person A would increase when hearing *ni*, resulting in the equal proportion of looks to Person A between the *ni* and *de* trials. Note that the looks to Person A were expected to be lower when hearing *ni* than *de* even for most of the L2 speakers, which made it possible to particularly assess the predictive ability of using the locative particles. This is because advanced Japanese L2 speakers with L1 Chinese were found to possess the knowledge of the destination usage of *ni* (see Hasuike, 2012).



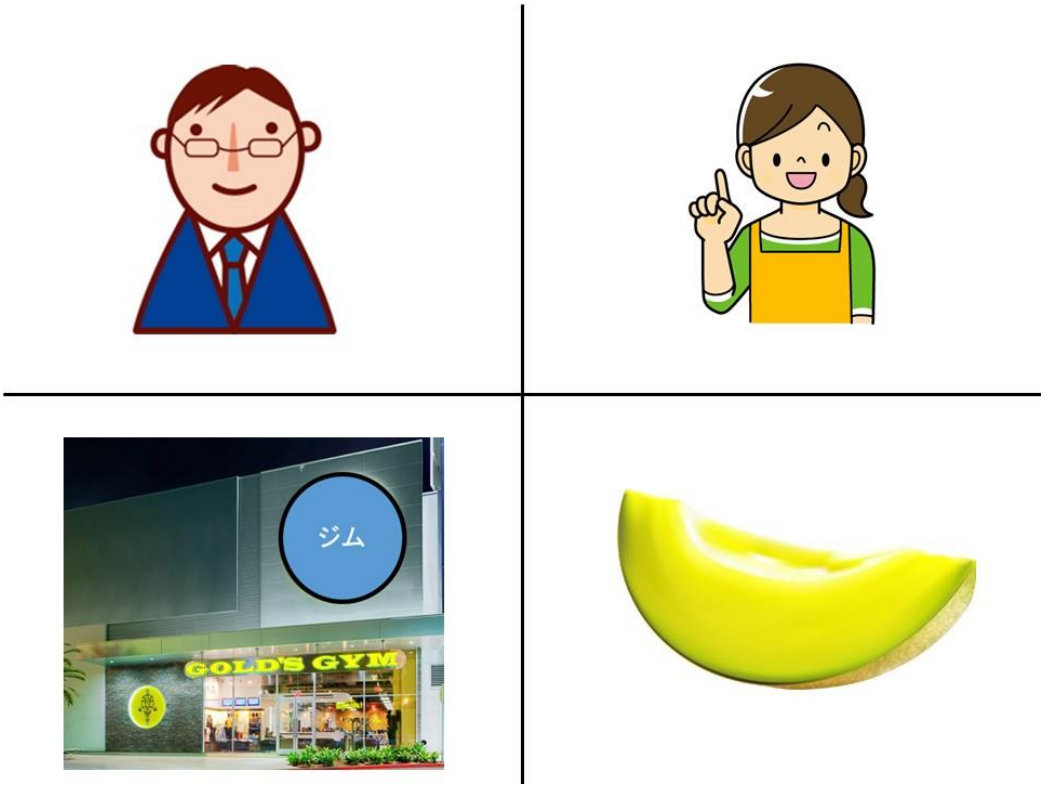
*Action (De): Otoko no hito ga densha **de katte ni** denwa wo shite imasu.*
Man SUB train-LOC selfishly calling is
(The man is calling selfishly on the train.)
*Direction (Ni): Otoko no hito ga densha **ni katte ni** norimashita.*
Man SUB train-LOC selfishly got on
(The man got on the train selfishly.)

Figure 8. Visual Scene and Critical Sentence for Ni/De

5.3.1.4 Tameni/Youni. To assess the acquisition of the constructions indicating purpose (*tameni/youni*), a visual scene was created to include Person A, Person B, and two objects (Figure 9). Person A (i.e., a father) was mentioned in the adverbial clause for the *tameni/youni* sentences, whereas Person B (i.e., mother) was only mentioned in the main clause with *youni*. Person A and person B were related in meaning (e.g., mother and father). Since the *youni* in the adverbial clause implicated that the subject of the adverbial clause and the one in the main clause were different, the proportion of looks to Person B is of interest when participants hear *youni*. In

contrast, the same subject was assumed to be involved in the main clause followed by the adverbial clause containing *tameni* because *tameni* could only permit the same subject between the two clauses. If participants were sensitive to the distinction of *tameni* and *youni*, then the looks to Person B would be greater in the *youni* sentence than in the *tameni* sentence.

The critical sentences in the two conditions were identical except for *tameni/youni*. The adverbial clause always started with NP (Person A) and ends with *tameni* or *youni*. As shown in the sample sentence in Figure 9, the father was the doer of the action in both sentences, but the person who went to buy a jersey was himself in the case of *tameni* and it was someone else (i.e., mother) in the case of *youni*. The region of interest was from the onset of *tameni/youni* to the offset of the object in the main clause (e.g., *jaaji o*), which corresponded to the onset of Person B in the *youni* sentences.



Tameni:

Otosan ga jimu de undousuru ***tameni, jaaji wo*** *kai ni ikimashita.*
Father-SUB gym-LOC do exercise PURPOSE, jersey-OBJ go buy.
(In order to do exercise at the gym, the father went out to buy a jersey.)
Youni:
Otosan ga jimu de undousuru ***youni, jaaji wo*** *okaasan-ga kai ni ikimashita.*
Father-SUB gym-LOC do exercise PURPOSE, jersey-OBJ mother-SUB go buy.
(So that the father does exercise, the mother went out to buy a jersey.)

Figure 9. Visual Scene and Critical Sentence for *Tameni/Youni*

5.3.2 Word-Monitoring Task. The word-monitoring task was used to measure the online sensitivity to grammatical errors. In this task, participants were instructed to listen to a spoken sentence for a target word and to press a button as soon as they identified it in the sentence. The target word appeared after the relevant target structure in a sentence; the difference in the Reaction Time (RT) to the target word between grammatical and ungrammatical sentences provided the index for the online grammatical sensitivity.

In this procedure, participants were first presented with the target word in the center of the screen. They were told to press the keyboard button as soon as they heard the target word in the sentence. The target word remained on the screen until the response was made. A yes/no comprehension question appeared on the screen, so that participants' attention was directed to the sentence meaning as well as the target word. This dual-task paradigm minimizes the application of explicit knowledge and strategy use.

The stimulus sentences for the four target linguistic structures are presented along with the target word in Table 7. For the transitive structures, both transitive and intransitive sentences were tested in grammatical and ungrammatical conditions. The target sentence always included a segment of the case-marking particle (*ga* or *o*) and a verb (transitive or intransitive). The target word was always the verb following the particles (*ga* or *o*). The particles (i.e., *o* and *ga*) were manipulated, rather than changing the transitivity of the verbs, because particles are less salient, and it is harder for participants to detect their ungrammaticality.

For classifiers, the classifier-noun pairs were embedded in a carrier sentence. For each of the eight classifiers, two nouns were chosen. The target word was the noun that followed the classifier and the genitive particle *no*.

For the *ni/de* construction, a segment of a location noun, the particle (*de/*ni*), and the action verb were always included in the sentence.[24] There are a limited number of stative verbs, so action verbs are chosen to test the sensitivity to incorrect usage of *ni*. The target monitoring word was the verb following *ni* or *de* (i.e., *neru*).

For the *youni/tameni* distinction, sentences started with an adverbial clause including either *youni* or *tameni* followed by the main clause. Subjects of the adverbial clause and the main clause were always explicitly stated, and the sentence with the *tameni* was ungrammatical. The target word was the subject in the main clause.

A list of stimulus sentences included 64 target sentences (16 for each structure) and 32 grammatical filler sentences. All the sentences were different from the ones used in the visual-world task. The lists were counter-balanced; half of the target sentences were grammatical, the other half were ungrammatical in List 1; the grammaticality of the sentences was reversed in List 2 so that no target sentences were presented twice to one participant. Half of the items (k = 48) were followed by a yes/no comprehension question. Not all the items were followed by a comprehension question in order to decrease the fatigue due to the longer time it takes to complete the task (see a similar approach in Jiang et al., 2011). The ratio was kept equal between

[24] Due to the constraints in the design of the visual-world task, the comparison of *ni* and *de* in the word-monitoring task and the self-paced reading task is slightly different from that in the visual-world task. In the visual-world task, the looks to the picture involving action (compatible with *de*) and those to the picture involving motion (compatible with *ni*) were compared. In the word-monitoring and self-paced reading tasks, the reaction time to the action verb (compatible with *de*) and that to the static verb (compatible with *ni*) were compared. The baseline is different (*ni* motion and static verbs), but the critical target structure (*de* for action verbs) is the same across the tasks.

a positive response and a negative response. The RT to the filler sentences was not analyzed. Before the actual test started, eight grammatical sentences were given for practice to familiarize participants with the task.

In sum, the same rationale applied to all four structures for measuring the online sensitivity to grammatical violations: If participants could use the information of the linguistic structures in real time (i.e., on the basis of implicit knowledge), RTs to the target word in ungrammatical sentences would be expected to be slower than that to the target word in grammatical sentences. The RT differences scores were computed by subtracting the RT to the grammatical sentences from the one to the ungrammatical sentences, which indexed the acquisition for each linguistic structure.

Table 7. *Sample Stimulus Sentences for the Word-Monitoring Task*

Condition	Stimulus Sentences
Transitive	*Ao to kiiro no enogu o/*ga **mazeru** to, kirei na mimdori ni naru.*
	Blue and yellow paint-ACC/SUB mix if, beautiful green become
	When you mix blue and yellow paints, it becomes beautiful green.
Intransitive	*Ao to kiiro no enogu ga/*o **mazaru** to, kirei na mimdori ni naru.*
	Blue and yellow paint-ACC/SUB mix if, beautiful green become
	When you mix blue and yellow paints, it becomes beautiful green.
Classifier	*Kumiko-san-wa san-dai/*satsu no **keitaidenwa** o motte imasu.*
	Kumiko-TOPIC three-CLASSIFIER GEN cell phone-OBJ have is
	Kumiko has three cellphones.
Ni/De	*Atatakai toki ni soto de/*ni **neru** to kimochi ii.*
	Warm time outside-LOC sleep if comfortable
	It is comfortable to sleep outside in the warm weather.
Purpose	*Musuko ga benkyou suru youni/*tameni, **hahaoya** ga atarashii tsukue wo katte ageta.*
	Son-SUB study PURPOSE, mother-SUB new desk-OBJ bought for him
	So that the son studies, his mother bought a new desk for him.

Note. The monitoring words are bolded and underlined.

5.3.3 Self-Paced Reading Task. Similarly to the word-monitoring task, the self-paced reading task assessed the online grammatical sensitivity while participants were reading the

sentence for comprehension. In the task, participants were asked to read a sentence word by word as quickly as possible and while paying attention to meaning to answer a comprehension question accurately. The first word of a sentence appeared on the left side of the screen, and when the button was pressed, the next word appeared to the right of the preceding word, which disappeared upon the presentation of the following word (moving-window presentation). When participants read the final word followed by the period, they pressed a second key to continue to either the next test item or a comprehension question.

As in Jiang (2007), the region of interest where RTs were compared between grammatical and ungrammatical sentences was at three word positions (see the underlined words in Table 8): at the critical word where the error occurred in the ungrammatical sentences (Region 1) and at the two words immediately following the critical word to capture spillover effects (Regions 2 and 3). The word preceding the critical region (i.e., Region 0) was also used as a baseline for checking that the reading time of the word before the critical region did not differ between grammatical and ungrammatical sentences. The first critical word was located in the same position as that in the word-monitoring task so that the effects could be compared fairly between the word-monitoring task and the self-paced reading task. If participants were sensitive to the grammatical error, which preceded the critical region, then their reading time would be delayed at (some of) these three positions.

As in the word-monitoring task, a list of stimulus sentences included 64 target sentences (16 for each structure) and 32 grammatical filler sentences. All the sentences were different from the other tasks. The lists were counter-balanced; half the target sentences were grammatical, the other half were ungrammatical in List 1; the grammaticality of the sentences was reversed in List 2 so that no target sentences were presented twice to one participant. Half of the items (k = 48)

were followed by a yes/no comprehension question (see a similar approach in Jiang et al., 2011). The ratio was kept equal between a positive response and a negative response. The RT to the filler sentences was not analyzed. Before the actual test started, eight grammatical sentences were given for practice to familiarize participants with the task.

In a similar way to the word-monitoring task, the index of linguistic knowledge was computed by subtracting the RTs for grammatical sentences from the RTs for ungrammatical sentences. There were three critical reading regions in the self-paced reading, and the first two regions were combined to compute the index to capture the earliest sensitivity to errors (see Analysis section).

Table 8. *Sample Stimulus Sentences for the Self-Paced Reading Task*

Condition	Stimulus Sentences
Transitive	*Gyunyuu to/ Chiizu to/ tamago o(*ga)/* ***mazetara,/ furaipan ni/ yasai to****/ issho ni/ iremasu.* Milk and/ Cheese and egg-OBJ(SUB) mix if frying pan-to vegetable together put into After you mix cheese and egg, you put them into the pan with vegetables.
Intransitive	*Gyunyuu to/ Chiizu to/ tamago ga(*o)/* ***mazattara,/ furaipan ni/ yasai to****/ issho ni/ iremasu.* Milk and/ Cheese and egg-SUB(OBJ) mix if, frying pan-to vegetable together put into After cheese and egg are mixed, you put them into the pan with vegetables.
Classifier	*Kono/ omise de/ juu-dai (*mai)-no/* ***keitaidenwa ga/ nusumareta to/ keisatsu ga****/ houkoku o/ uketa soudesu.* This shop-LOC ten-CLASSIFER GEN cell phone-SUB stolen that policie-SUB report-OBJ received heard I heard that the police received the report that ten cellphones were stolen at this shop.
Ni/De	*Kinou ane to/ ohiru no / san-ji goro/ resutoran de (*ni)/* ***shokuji shinagara,/ ryokou no/ hanashi o****/ shimashita.* Yesterday sister-with noon-GEN three about restaurant-LOC eat while trip-GEN talk did I talked about a trip with my elder sister at the restaurant [at] about 3 in the afternoon.
Purpose	*Jibun no kodomo ga/ takusan/ gohan o/ taberu youni(*tameni)/* ***chichioya ga/ oishii/ ryouri o/*** *tsukutte/ ageta.* My child-SUB a lot of food-OBJ eat PURPOSE father-SUB good dishes-OBJ cooked for him. So that his child eats a lot, the father cooked good dishes.

Note. The critical regions are bolded and underlined.

5.3.4 Timed Auditory GJT. In the computer-delivered timed auditory GJT, participants listened to an aural stimulus sentence and indicated whether each sentence was grammatical or ungrammatical by pressing a response button. They were asked to press a key as soon as an error was detected in the sentence. Previous studies like R. Ellis (2005) and Bowles (2011) set the time limit for each sentence based on the NSs' average response time plus an additional 20% of the time for each sentence. This 20% criterion might be too short; some participants were likely to be discouraged to perform the task after not being able to respond to many of the items within the short time limit. This was particularly considered to be of concern because the three similar time-pressured tests were to be administered in a row, which might discourage participants from taking the test seriously. Instead of following the exact same procedure as the previous studies, a more lenient time pressure was imposed on the tasks—10 seconds across all the test items. In addition, responses that were made longer than the time limit were dealt with after administering the test (see Analysis for details).

The stimulus sentences consisted of 64 target sentences (16 for each structure), half grammatical and half ungrammatical. The sentences were different from the ones used for the other tasks. No filler sentences were included. As the responses with long RTs were to be omitted from analysis, the instructions urged participants to respond as quickly as possible. Before the actual test, participants took a practice session. During the practice phase, participants were presented with four practice sentences, two grammatical and two ungrammatical, to familiarize them with the task. After responding to each of these practice sentences, participants were presented with a reminder for three seconds, saying "Please respond quickly based on your

intuition." No feedback or remainder were given during the actual test phase. The percentage scores for accuracy were calculated separately for grammatical and ungrammatical items and for both types of items combined.

5.3.5 Timed Written GJT. As in the timed auditory GJT, the timed visual GJT was also administered on a computer. The procedure was identical to the one in the timed auditory GJT except for the modality of the stimulus sentences. Participants were presented with a written sentence on a screen and asked to indicate whether each sentence was grammatical or ungrammatical by pressing a response button as quickly as they could. They were allowed to press the key while the sentence was played when the error was detected within the sentence. The time limit imposed on the task was 10 seconds across all the items.

The stimulus sentences consisted of 64 target sentences (16 for each structure), half grammatical and half ungrammatical. They were different from the sentences in the other tasks. No filler sentences were included. The same practice session as the auditory GJT was given before the actual test. The percentage accuracy score was calculated for all the items and separately for grammatical and ungrammatical items.

5.3.6 Timed Fill-in-the-Blank Test (SPOT). In a timed fill-in-the-blank test (SPOT[25]), the participants were presented with a single sentence with some blanks on the computer screen. Then, they had to fill in the blank with Japanese characters on the answer sheet as quickly as they could. A blank in each of the sentences was made to specifically target one of the linguistic structures in the current study. Once they filled out the answer on the sheet, they pressed a computer button to move on to the next item. Although participants were told to respond as

[25] Since this procedure was similar to the format of existing tests in the Japanese education system, where it is called the Simple-Oriented Performance Test (SPOT), this task is called the timed SPOT here (Kobayashi, Ford-Niwa, & Yamoto, 1996).

quickly as they could. Due to the fact that the experimenter accidentally set the time limit to 100 seconds instead of 10 seconds, the time limit was virtually not imposed accidentally. The experimenter set the maximum time allowed to 100 seconds; most L2 speakers did not dwell on the items for more than a few seconds, and the same stringent cut-off was applied after the test administration as in the other two GJTs, however.[26]

Since the Japanese language uses three different kinds of written scripts (Chinese characters, *hiragana,* and *katakana*), only syllabic characters, *hiragana*, were used to fill in the blanks. The number of characters required was indicated by the number of blank circles in the sentence. For example, the following four sentences illustrate how stimulus sentences for each of the four target structures were tested. For instance, the blank circle in the first example should be filled with the object marking particle *wo* (を), whereas the second sentence has two circles and they should be filled in with the two hiragana letters, *ken* (けん). The target structure of sentence (4) does not have circles, but offers a choice between *tameni* and *youni* instead. The answer options were given because the pilot study showed that several NS participants performed poorly when answer options were not given (see also the results of *tameni/youni* by NSs).

(1) Transitive/Intransitive
パスワードを知らないまま金庫〇閉めたので、金庫から宝石が取り出せません。
Password wo shiranai mama kinko 〇 shimeta node, kinko kara houseki ga toridasemasen.
Password-OBJ without knowing safe-OBJ close because safe from jewel-SBU can't take out
(2) Classifiers
この町の消防署の隣には、２〇〇のコンビニが並んで建っている。
Kono machi no shoubousho no tonari niwa ni-〇〇 no konbini ga narande tatte iru.
This town-GEN fire station-GEN beside 2-Classisifier GEN convenience store-SUB built
(3) *Ni/De*
多くの高校生は友達とカラオケボックス〇歌って踊ります。
Ooku no koukousei wa tomodachi to karaokebokkusu 〇 utatte odorimasu.

[26] The average response time by L2 spekers was 9.00 seconds (SD = 3.07, range = 3.41-17.39).

Many high schoolers-SUB friends with karaoke box-LOC sing and dance

(4) *Tameni/Youni*

加藤さんが英語だけを話す(ため/よう)に、英語の先生が英語で話しかけます。

Kato-san ga eigo dake wo hanasu (tame/you)ni, eigo no sensei ga eigo de hanashikakemasu.

Mr. Kato-SUB English only OBJ speak PURPOSE English teacher-SUB English speak

The stimulus set consisted of 64 target sentences (16 for each structure); they were different from the sentences in the other tasks. No filler sentences were included.

5.3.7 Explicit Learning Aptitude: LLAMA F. An adaptive version of LLAMA F (Meara, 2005) was administered to measure inductive language learning ability. Explicit learning aptitude was operationalized as the score on LLAMA F. It was defined as "the ability to infer or induce the rules governing a set of language materials, given samples of language materials that permit such inferences" (Carroll, 1981, p. 105). LLAMA F consisted of a learning phase and a test phase. In the learning phase, participants were given five minutes to learn a new language by seeing sentences matched with pictures. In the testing phase, the program displayed a picture and two sentences, one grammatical and the other ungrammatical, and their task was to choose the grammatical sentence. The test consisted of 30 items. Participants were required to induce the rules of grammar by looking at pictures and word sequences; in this way, LLAMA F assessed language analysis ability independently from the participants' first language.

5.3.8 Implicit Learning Aptitude: SRT task. The Serial Reaction Time (SRT) task was administered to measure aptitude for implicit sequence learning. Implicit learning aptitude was operationalized as the domain-general ability to learn sequences without awareness. The probabilistic SRT task adopted from Kaufman et al. (2010) was used in the present study. In the SRT task, participants saw a dot, appearing at one of four locations on the computer screen, and responded to it as quickly and accurately as possible by pressing the corresponding key. Unbeknownst to participants, the sequence of stimuli was generated by a probabilistic rule: 85%

of the sequences follow the rule (probable, training condition), whereas the other 15% of the sequence were generated by another rule (improbable, the control condition). More specifically, Sequence A (1–2–1–4–3–2–4–1–3–4–2–3) occurred with a probability of 0.85, and Sequence B (3–2–3–4–1–2–4–3–1–4–2–1) occurred with a probability of 0.15 in one block. This probabilistic nature of the SRT task made it difficult to learn the sequence explicitly. It is noted that these sequences were comprised entirely of second-order conditionals, so they could not be determined by first-order conditionals (Reed & Johnson, 1994); a second-order conditional sequence was determined by the previous two locations, not by the previous location, which made the task more complex and minimized chunk learning. There were eight blocks, and each block consists of 120 trials, 960 trials in total. The SRT task was scored by subtracting the mean RTs in the training condition (Sequence A) from those in the control condition (Sequence B), which reflected the amount of learning.

After the SRT task, participants also took a surprise recognition test with confidence ratings. The test assessed whether participants became aware of the sequence patterns in the SRT task, i.e., whether they developed explicit knowledge about the sequence (Granena, 2013b; Shanks & Johnstone, 1999). It consisted of 24 triads, half old (familiar) and half new (less familiar). Following Granena (2013b), the 12 old triads were constructed following second order conditionals in Sequence A (3–4–2, 3–1–2, 1–4–3, 2–4–1, 4–2–3, 1–2–1, 4–3–2, 4–1–3, 2–3–1, 2–1–4, 3–2–4, 1–3–4) and the 12 novel ones were constructed following second order conditionals in Sequence B (3–4–1, 3–1–4, 1–4–2, 2–4–3, 4–2–1, 1–2–4, 4–3–1, 4–1–2, 2–3–4, 2–1–3, 3–2–3, 1–3–2). The first two locations in every triad were the same in both old (Sequence A) and new (Sequence B) triads, but the third location was different (e.g., transition 3–4 was followed by location 2 in Sequence A and by location 1 in Sequence B). Because participants

were trained to learn the second-order conditional information of Sequence A, they were expected to respond to the third location faster in the old triads. After every test item, participants rated their familiarity by giving a confidence rating on a 6-point scale ranging from 1 (I'm sure that this sequence was part of the test) to 6 (I'm sure that this sequence was not part of the test). Evidence of poor recognition, but faster RTs, for segments of the old sequences was taken to suggest that the knowledge acquired during the training task produced behavioral effects before these effects are consciously attributed to the results of learning—implicit knowledge.

5.3.9 Phonological Short-Term Memory: Letter Span Task. The letter-span test in the Hi-LAB was used to measure phonological short-term memory (Linck, Hughes, et al., 2013). This aptitude component was measured to examine whether memory for verbal information could moderate the effects of explicit and implicit learning aptitudes. In this task, a list of letters was presented on the screen at 900 ms intervals, and participants were asked to recall the letters in order. The length of the list varied from three to nine letters, and for each of the 7 lengths, three lists were presented, for a total of 21 lists, in pseudorandom order. The letters were drawn from a set of 12 consonants. The score was based on the total number of letters recalled in their correct positions.

5.3.10 Summary. The current study employed a within-subject design in which all participants took all the linguistic knowledge tests. Given the primary aim of this study to validate explicit/implicit knowledge measures, the possibility that the target structures were identified by participants had to be minimized; the language tasks were designed to minimize that possibility.

First, no identical sentences appeared across the five tests. Second, since the same target transitive/intransitive verbs and the same classifier-noun pairs could appear across the tasks, the

target words were counter-balanced across the tasks (see Appendix C and Appendix D). In the transitive structure and the classifier structure, four lists were created across the four linguistic tasks that utilized both grammatical and ungrammatical sentences (i.e., the word-monitoring task, the self-paced reading task, the auditory GJT and the written GJT), so that each list uses the same verb or classifier-noun pair in grammatical and ungrammatical conditions only in one of the tasks. In other words, the words that appear in the grammatical condition in one task never appeared in the other tasks as a grammatical condition. Third, the ratio of target sentences was decreased in the implicit knowledge tests. The visual-world task, which lies at the far implicit end of the continuum of explicit to implicit knowledge, did not include any ungrammatical sentences; and the target sentence was always followed by a filler sentence (see Table 9). In the word-monitoring task and the self-paced reading task, 32 grammatical fillers were included to reduce the ratio of ungrammatical sentences to 33%. No fillers were included in the two timed GJTs because they were form-focused tasks. Having no fillers could also reduce the time and the effects of fatigue when completing the task.

Table 9. *Ratio of Grammatical and Ungrammatical Target Sentences and Fillers*

	Target (Gramm.)	Target (Ungramm.)	Filler (Gramm.)	Ratio Gramm.
Visual World	64	0	64	100%
Word Monit.	32	32	32	67%
Self-paced R.	32	32	32	67%
Timed AGJT	32	32	0	50%
Timed VGJT	32	32	0	50%
Timed SPOT	64	0	0	100%

5.4 Procedure

Participants were tested individually in a soundproof booth, with the nine tests in the fixed order described in Table 10. After the consent form and the background questionnaire, the linguistic tasks were administered first from the most implicit linguistic tasks to the more

explicit. Individual difference measures were administered after the linguistic test battery. A three-minute break was interspersed between the tasks. Participants were also provided with snacks and drinks and allowed to take a break between the tests as needed. The whole session approximately took 2.5 to 3 hours.

Table 10. *Order of Tests*

Tasks	Min.
1. Consent Form & Background Questionnaire	15
2. Visual-World Task	30
Break	3
3. Word-monitoring task	20
4. Self-paced reading task	20
Break	3
5. Timed auditory GJT	10
6. Timed visual GJT	10
7. Fill-in-the-blank	10
Break	3
8. SRT task + Recognition Test	20
9. Letter Span task	10
10. LLAMA F	15

5.5 Data Analysis

The primary goals of the present dissertation were (1) to develop valid behavioral measures for automatized explicit knowledge and implicit knowledge and (2) to investigate the interface issue of explicit and implicit knowledge through the role of cognitive aptitudes. Before discussing these two main analyses, scoring of each task will be explained. The statistical approach for the construct validation of the measures will then be delineated, followed by the analysis for the role of aptitudes as predicting variables.

5.5.1 Visual-World Task. To ensure that the participants were focusing on meaning to be able to answer the comprehension questions, accuracy of the comprehension questions was

computed. The mean accuracy scores were 93.48 (*SD* = 2.06) for the NSs and 90.83 (*SD* = 4.12) for the L2 speakers. In previous self-paced reading research, which has a similar design to a visual-world task and a word-monitoring task, a participant whose error rate was higher than 25% was excluded from analysis of dependent variables (e.g., RT) to ensure that each individual was paying attention to meaning (Jiang et al., 2011). None of the participants scored below 75% for the visual-world task; all participants' eye-movements data were analyzed.

Location of fixation was coded as a look towards one of the quadrants or missing due to blinking or looks outside of the screen. The missing frames accounted for 5.45% of the NSs' data and 5.87% in the L2 speakers' data. Mean proportion of fixations over trials was plotted separately for each target structure and groups. The fixations in plots were time-locked from the onset of the target linguistic triggers (i.e., disambiguation point) to the end of critical region plus some buffer. Each period was shifted 200ms after the trigger of linguistic cues in speech to account for the time it took, physiologically, to generate saccadic eye-movement (Matin, et al., 1993).

Although the fixations were first time-locked to the onset of the target linguistic trigger, it was unknown exactly when listeners started to use the cues from each of the four target grammatical structures. Depending on linguistic structures, the time it would take to deploy linguistic information might be different. In order to identify the data-driven onset of the fixations that were triggered by each of the target structures, fixation plots were inspected to find the time point where fixations in one condition differentiated from the other, by inspecting the standard error bars of mean. The data-driven onset was defined as the time window in which the mean Target advantage plus one standard error in the Target trials became greater than that in the non-Target trials. Once the data-driven onset of fixations was identified, the post-hoc region was

determined from the data-driven onset to the end of the critical region. For this post-hoc region, paired-samples t-tests were conducted to statistically examine whether the looks in one condition were significantly higher than the other. The post-hoc region was set based on the native speakers' data, not on L2 speakers. It was assumed that L2 speakers were slower to show the effects of linguistic cues. The same criteria as for the NSs were used, however, because the current study attempted to assess L2 implicit knowledge that were qualitatively similar to NSs.

5.5.2 Word-Monitoring Task. Comprehension accuracy scores were computed to check whether the participants were focusing on meaning to perform the task. The mean accuracy scores were 95.92 (*SD* = 2.48) for the NSs and 91.55 (*SD* = 5.14) for the L2 speakers. None of the native speakers or L2 speakers had an error rate above 25%; all the participants' RT data were analyzed. In order to screen the RT data, outliers were discarded that fell outside the low and high cutoffs set at 100 ms and 2500 ms or that were 3 SDs above or below each participant's mean, respectively. The higher cutoff was set in order to exclude responses in which participants inadvertently forgot to respond to the target word, and the lower cutoff was set to exclude the responses given without hearing a target word. These procedures, along with display errors, eliminated 2.7% and 3.3% of the data for NSs and L2 speakers, respectively. Paired t-tests were conducted to compare the RTs for grammatical and ungrammatical items across the target structures. Split-half reliability with Spearman-Brown correction was .915 for RT data of L2 speakers.

5.5.3 Self-Paced Reading Task. As in the visual-world task and the word-monitoring task, the accuracy of the comprehension questions was examined first. The mean accuracy scores were 96.65 (*SD* = 2.46) for the NSs and 91.55 (*SD* = 4.82) for the L2 speakers. None of the native speakers or L2 speakers had an error rate above 25%; all the participants' RT data were

analyzed. In order to screen the RT data, outliers were discarded that fell outside the low and high cutoffs set at 120 ms and 1500 ms or that were 3 SDs above or below each participant's mean, respectively. The higher cutoff was set in order to exclude responses in which participants were reading too slowly, and the lower cutoff was set to exclude the responses given without reading the words. These procedures, along with display errors, eliminated 3.4% and 6.6% of the data for NSs and L2 speakers, respectively. Split-half reliability with Spearman-Brown correction was .978 for RT data of L2 speakers.

In order to show that the NSs performed the self-paced reading task appropriately, the RT differences were examined by paired-samples *t*-tests at each of the three critical regions. It is important to examine the sensitivity for each region because native speakers and L2 speakers might be different with respect to the region where they showed the sensitivity to grammatical errors. To illustrate how RTs were analyzed for each target structure, some examples from the transitive verb are given below. The underlined regions in the sentences (6a) and (6b) with the transitive verb illustrate the four regions where the mean RTs were analyzed:

6a. Gyunyuu to/ Chiizu to/ tamago o/ mazetara,/ furaipan ni/ yasai to/ issho ni/ iremasu.

*6b *Gyunyuu to/ Chiizu to/ tamago ga/ mazetara,/ furaipan ni/ yasai to/ issho ni/ iremasu.*

Region 0 Region 1 Region 2 Region 3

Milk and/ Cheese and egg-OBJ(SUB) mix if frying pan-to vegetable together put into

After you mix milk, cheese and egg, you put them into the pan with vegetables.

The first underlined region was the position before the error, Region 0. The error position was marked as Region 1, followed by the two spillover regions, Region 2 and Region 3. No

difference in RTs would be expected in Region 0 because the two versions of the sentence were identical up to that point. RT differences would be expected to occur at any or all of the following regions (Regions 1-3).

5.5.4 Setting the time limit on the Timed Form-Focused Tests.[27] Instead of imposing strict time pressure (e.g., 20% plus NSs' mean RTs) on the task, GJTs were given under less time pressure, but L2 speakers' responses were scored incorrect if the response time was not within a certain time limit based on the NSs' RTs. We first examined what percentage of the data was lost within the NSs by giving no credit to the responses whose response rate was above the mean RTs plus 20% for each test item. The percentages of the responses that were within the cut-off point varied across the three tests[28]: 84.74% in the auditory GJT, 68.82% in the visual GJT, 77.76% in the timed SPOT. Particularly when the tasks were administered in the visual modality, there was more variability in reading time, even among NSs. This is even more problematic for L2 speakers, as reading speed might affect whether they are able to use linguistic knowledge within the time limit. When this 20% cut-off point RTs was imposed on L2 speakers' data, the remaining cases were extremely scarce, particularly in the visual GJT: 37.09% in the auditory GJT, 8.66% in the visual GJT, and 18.37% in the SPOT. NSs' fast reading speed made the cutoff too strict for L2 speakers' data. We decided to first set a different percentage value so that fewer numbers of the NSs' responses were excluded. More specifically, given the variability across the three tests, percentages to be added to the NSs' mean RT were determined such that the NSs' mean error rate of the total score was kept less than 10%. In other words, we identified the cutoff percentages that kept 90% of the NSs' responses; this covered a wider range of L2 speakers'

[27] One item in the auditory GJT and another in the visual GJT were excluded from the analyses due to the low accuracy rate for NSs (the scores were 58% and 68%, respectively).
[28] The items of *tameni*/*youni* were excluded from these analyses, as NSs did not show sensitivity in the task.

responses. The cutoff percentages that retained 90% of NSs data were mean RTs + 50% for the auditory GJT, mean RTs + 120% for the visual GJT, and mean RTs + 50% for the SPOT. How these more lenient cut-off points affected the responses of L2 speakers in the three tests will be presented below. Note that we admit there should be different ways of setting the cutoff, and the different procedure made it more difficult to compare the GJT results from previous studies, but this post-hoc data trimming procedure was chosen in order to avoid a situation in which L2 speakers lost the motivation to perform the tests.

5.5.4.1 Auditory GJT. As delineated above, NSs' RTs were used as a baseline to determine the cutoff for each of the test items. In order to retain the 90% of NSs' responses, 50% was added to the mean RT for each individual item, and the cutoff range was from 2.10 seconds to 8.16 seconds. When this 50% cutoff was applied to L2 speakers' data, 37.38% of the data were excluded from further analysis.

5.5.4.2 Visual GJT. In order retain 90% of the NSs' responses, the cutoff for the visual GJT was 120% plus the mean NSs' RTs. 120% was added to the mean RT for each individual item, and the cutoff range was from 3.08 to 17.07 seconds.[29] When this cutoff was applied to L2 speakers' data, 52.94% of the data were excluded from further analysis.

5.5.4.3 SPOT. In order retain 90% of the NSs' responses, the cutoff for the SPOT was 50% plus the mean NSs' RTs. An extra 50% was added to the mean RT for each individual item, and the cutoff range was from 3.47 to 10.93 seconds. When this cutoff was applied to L2 speakers' data, 67.63% of the data were excluded from the further analyses.

5.5.5 Construct Validation of Explicit and Implicit Knowledge Measures. In the analysis for construct validation of explicit and implicit knowledge measures, three statistical

[29] There were five test items whose cutoffs were longer than 10 seconds, but as the time limit on the test was 10 seconds, they were automatically scored incorrect.

approaches were employed: (1) Confirmatory Factor Analysis (CFA), (2) Multi-Trait Multi-Method (MTMM) analysis, and Structural Equation Modeling (SEM). The present study hypothesized that the three online-sentence processing measurements (the visual-world task, the word-monitoring task, and the self-paced reading task) tap into implicit knowledge, whereas the three form-focused tests (the timed auditory GJT, the timed visual GJT, the timed SPOT) tap into automatized explicit knowledge. In total, six variables were submitted to the analysis. The hypotheses were tested through the sophisticated statistical analyses above because these analyses could account for the relationship among all the measures at the same time, diminishing spurious relationships (which could be found in zero-order correlations), and could assess the measurement errors.

Since the previous study by Suzuki and DeKeyser (in press) suggested that the amount of exposure in L2, approximately indexed by the length of residence (LOR) in Japan, this might influence the type of linguistic knowledge L2 speakers have recourse to in a given task. The CFA models were estimated separately for the two LOR groups of L2 speakers. The whole group was split into half by using the median LOR, 39 months (see Table 11). According to independent t-tests, the two groups were significantly different in terms of length of residence and age at testing ($p < .001$). The other factors (age of arrival, onset of instruction and length of instruction) were not different ($p > .1$). Since each model consists of six indicators, even the rough estimation of the necessary sample size (10 participants * 6 indicators = 60) indicates the sample size was less than ideal. The results from the subset analyses should be interpreted cautiously.

Table 11. *Background Information for Short-LOR and Long-LOR Groups*

	Age at Testing	Age of Arrival	Length of Residence (months)	Onset of Instruction	Length of Instruction (months)
Short-LOR (n =48)					
Mean	23.88	21.21	30.13	18.69	41.54
SD	2.72	2.63	4.33	1.82	17.16
Min	19	17	24	13	6
Max	32	29	38	24	72
Long-LOR (n =52)					
Mean	27.90	21.50	63.13	19.31	40.71
SD	4.91	2.72	30.66	2.57	17.84
Min	22	17	39	13	3
Max	47	30	197	27	84

5.5.5.1 Confirmatory Factor Analysis. A Confirmatory Factor Analysis (CFA) evaluated the extent to which hypothesized relationships between the measurements could be confirmed. Three models were constructed. Figure 3 presents the hypothesized best-fitting two-factor model: implicit knowledge (three online methods) and explicit knowledge (grammatical and ungrammatical components from the two GJTs). This model was motivated by the hypothesis proposed by Suzuki and DeKeyser (in press) that real-time anticipatory sentence processing while attention was directed to meaning could be an indicator of implicit knowledge, whereas time-pressured form-focused tasks should draw on automatized explicit knowledge (See Table 6).

The second, one-factor model stood in complete contrast with the first model (Figure 4): all the variables load on the single implicit knowledge factor. This model is motivated by the validation studies by R. Ellis (2005) and Bowles (2011), in which timed GJT was loaded on the implicit knowledge factor along with the other implicit knowledge measures (EI and the oral narrative task). A third model was constructed in terms of modality of measurements: a written/aural model (Figure 10).

For all the models, the errors first were correlated between the two similar methods: (1) the timed auditory GJT and the timed visual GJT and (2) the word-monitoring task and the self-paced reading task. After the model was fitted, only significant correlated errors were retained. It was predicted that the first model would fit the best among the three models.

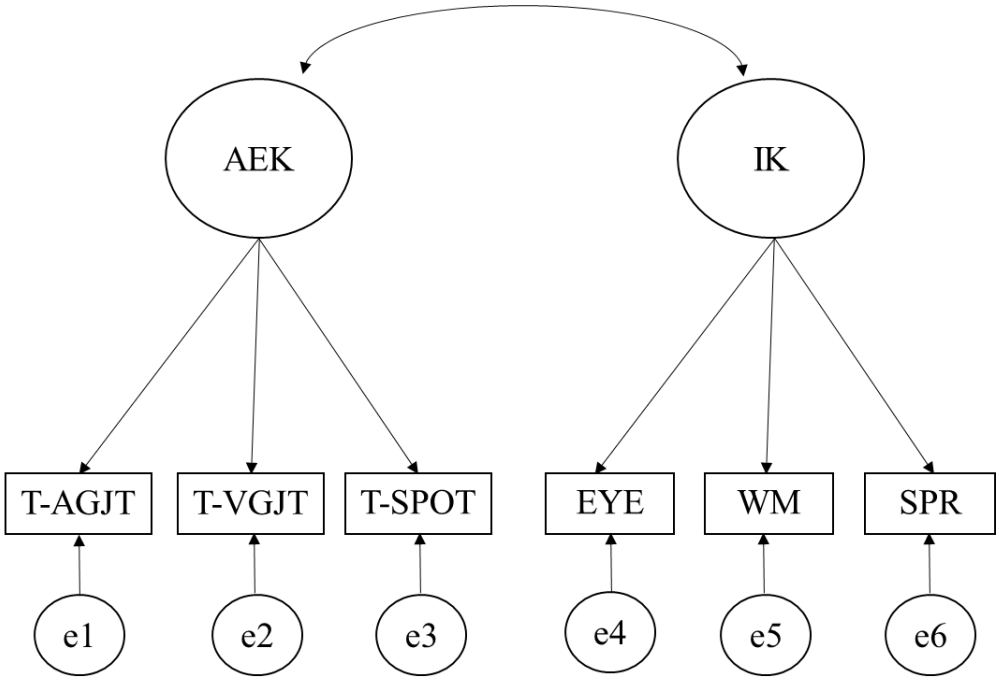


Figure 3. CFA Model 1: Two-Factor Model
Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

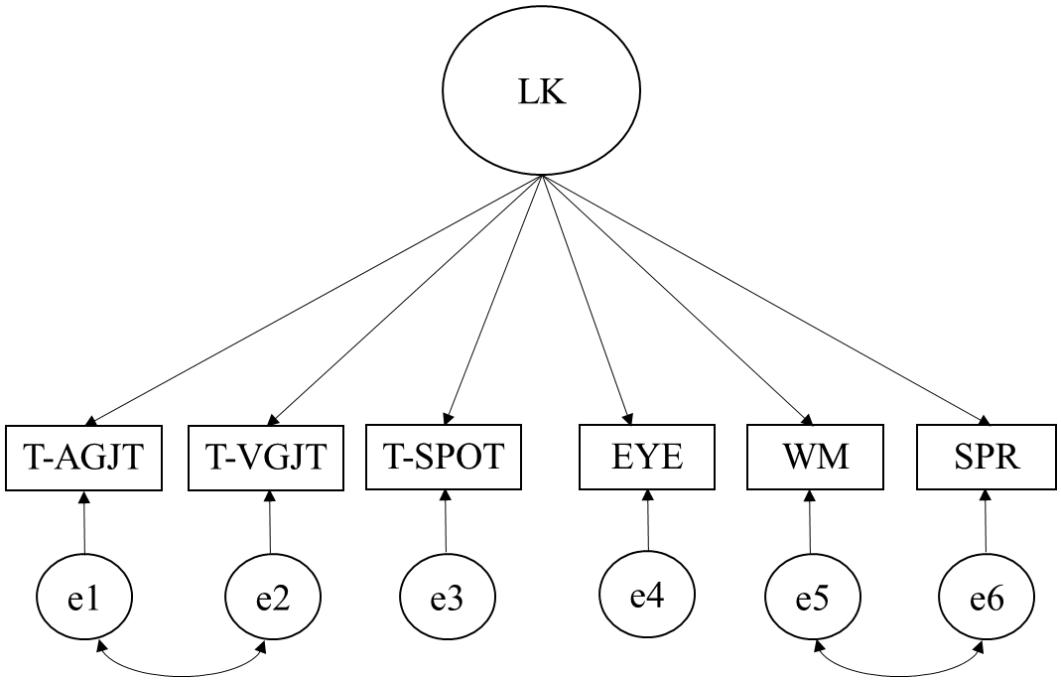


Figure 4. CFA Model 2: One-Factor Model
Note. LK = Linguistic Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

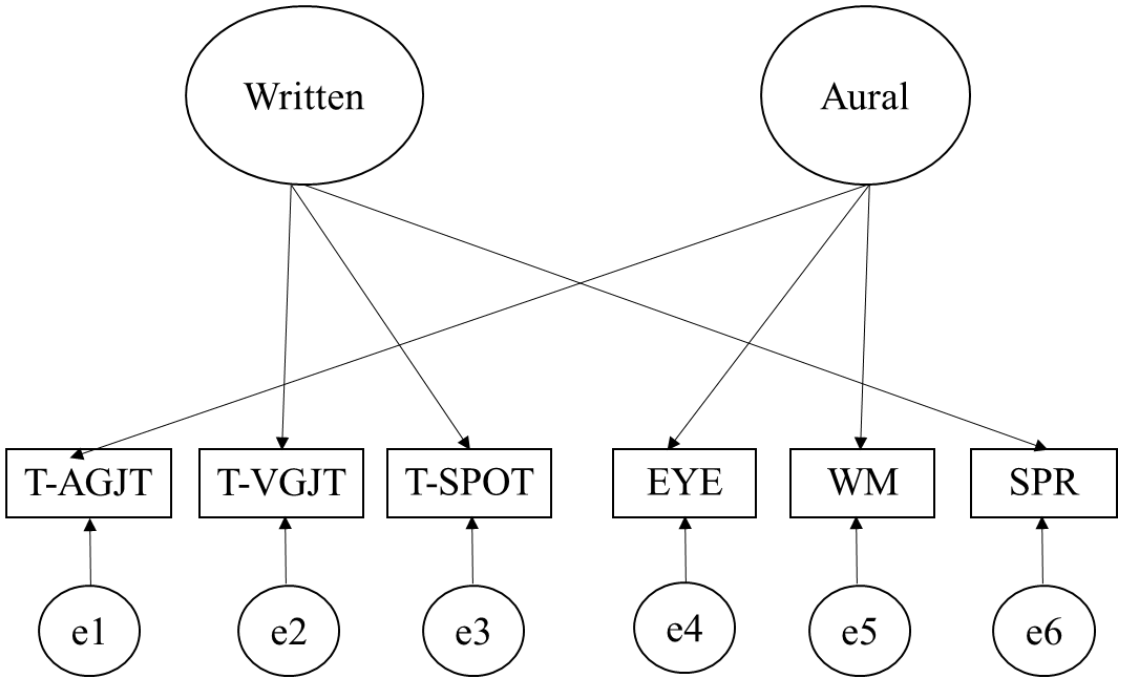


Figure 10. CFA Model 3: Written and Aural Model

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

In order to evaluate the models statistically (Hypothesis 1), a maximum-likelihood method was used to estimate the model parameters as implemented in the software package LISREL 9.1(Jöreskog & Sörbom, 2013). Since the chi-square statistics cannot be used as a sole indicator for model fit, multiple fit indices should be jointly used to assess the model fit (Brown, 2006; Hoyle & Panter, 1995). The following three categories of fit indices were utilized to assess the overall goodness of fit of the CFA models: (1) absolute fit indices (Standardized Root Mean Square (SRMR)), (2) incremental fit indices (the comparative fit index (CFI) and the Benter-Bonnet non-normed fit index (NNFI)), and (3) fit indices adjusting for model parsimony (Root Mean Square Error of Association (RMSEA) and Akaike Information Criterion (AIC)).

According to the findings of simulation studies conducted by Hu and Bentler (1999), a good fit between the target model and the observed data (ML estimation) was obtained in instances where (1) SRMR values were below .09; (2) RMSEA values were below .06; and (3) CFI and NNFI were above .96. Based on these empirically derived criteria, each of the models was assessed to exhibit either three levels of fit: good fit, marginal fit, and poor fit. When the indices in the two or three categories out of three met the criteria above, the model was considered to be a good fit (Hu & Bentler, 1999). When the index from one category met the criterion, the model was considered to be a marginal fit. When none of the fit indices reach the criteria, the model was considered to be a poor fit. After assessing the models individually, all

the (nested) models were compared directly by the goodness-of-fit testing indexed by the chi-square statistics.

In order to seek evidence for convergent validity (hypothesis 2), the magnitudes and the significance of the factor loadings were examined. A latent construct reliability was also assessed by a coefficient H) computed from the standardized factor loadings as in the formula below (Hancock & Mueller, 2001). Coefficient H can assess the stability of a construct as reflected in the data on the measurements, and it is based on the squared standardized loadings of individual measurements. The coefficient ranges from 0 (if all the standardized loadings are 0) to 1 (if a single standardized loading is 1 or -1); it can be interpreted that the value of .70 or higher indicates good reliability and the value between .6 and .7 are acceptable (Hair, Anderson, Tatham, & Black, 1998). High construct reliability indicates that the measures consistently represent the same latent construct.

$$H = \frac{1}{1 + \frac{1}{\frac{\ell_1^2}{(1-\ell_1^2)} + \ldots + \frac{\ell_p^2}{(1-\ell_p^2)}}}$$

Note: ℓ is a standardized factor loading, and p is the number of indicators of the factor of interest.

The divergent validity was assessed by the correlation between the two latent factors as well as the fit of the one-factor model (Hypotheses 3a and 3b).

5.5.5.2 Multi-trait Multi-method Analysis. In order to investigate the construct validity of measurements more rigorously, a CFA model of multi-trait-multi-method (MTMM) analysis was conducted to determine the extent to which variance in the measurements could be attributed to latent constructs of linguistic knowledge (traits) and to specific methods (hypothesis 4). In

other words, this analysis allowed for estimating the effects of methods or identifying the artifact of the relationship between the methods due to the same elicitation technique used.

Since the current study only has two similar methods, the factor loadings of indicators loading on the same trait factor are constrained to equality (Brown, 2006). There are two types of MTMM analysis: a correlated methods model and a correlated uniqueness model. Since a correlated methods model cannot be conducted to the design with two traits with two methods, a correlated uniqueness model was chosen (Brown, 2006). A two-trait and two-method model was fit to the data by setting the factor loadings of indicators loading on the same trait factor constrained to equality (Brown, 2006, p. 220). The correlated uniqueness model is presented in Figure 11, in which two traits and two methods were specified. Six measurements or variables were drawn in the rectangles between the traits and the methods. Two traits were specified (automatized explicit knowledge and implicit knowledge), which were drawn in the circles located in the upper side of the model. Two assessment methods, which were drawn in the lower side of the model, were represented by the correlated error between the visual GJT and the auditory GJT (grammaticality judgment) and the one between the word-monitoring task and the self-paced reading task (reaction time measurements).

Based on the hypotheses of the present study, the following predictions were presented regarding the path coefficients and covariance among the traits, measurements, and the methods: The trait factor loadings would be large and statistically significant (convergent validity). A weak and non-significant correlation between the two-trait factors would be expected (discriminant validity). Last, the method effects, indicated by the correlated errors between the similar methods, were expected to be smaller than the trait effects.

As in the CFA, the maximum likelihood method was used to estimate model parameters. Model fit was evaluated by the multiple indices.

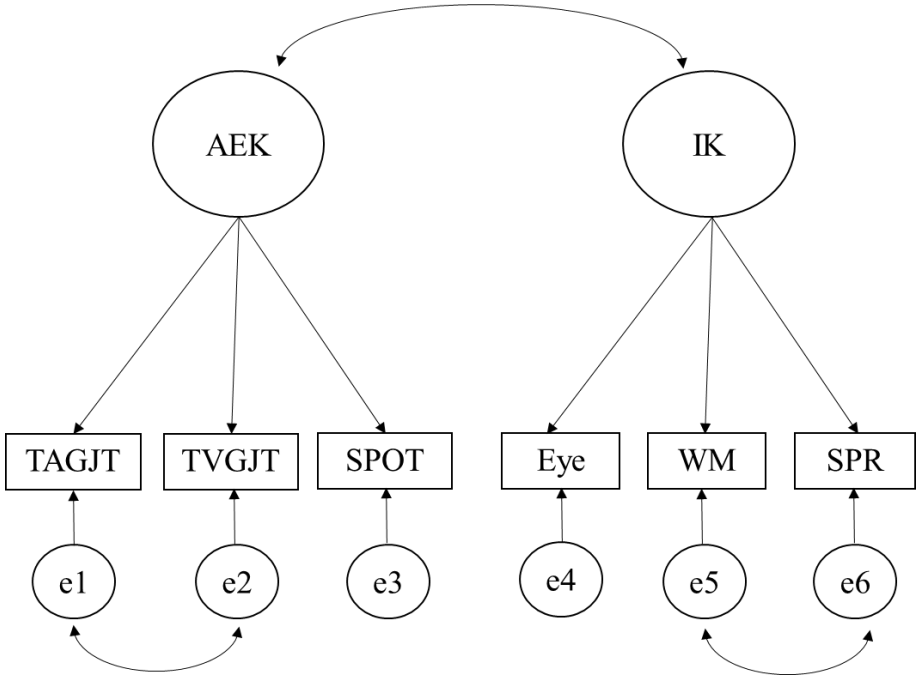


Figure 11. MTMM Model 1: Correlated Uniqueness model

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

5.5.5.3 Structural Equation Modeling. After the two hypothesized constructs were modeled both in the CFA and MTMM analyses, the two constructs were examined for whether they were really labeled as explicit and implicit knowledge. A SEM validation model was examined in Figure 12. Two latent factors of cognitive aptitudes were added to the best-predicted CFA model: implicit learning aptitude and explicit learning aptitude. The measurement errors were fixated on the indicators based on the reliability indices of the tests, respectively. The measurement errors were calculated by subtracting the reliability coefficients from 1: for the SRT task (*ME* = .48) and for the LLAMA F (*ME* = .10). By positing the latent factors, instead of measurement indicators alone, the accuracy of the model estimation improves because the

measurement errors were accounted for at the measurements levels.[30] As in the CFA models, the two correlated errors were added first and kept only if they were significant.

It was predicted that the path loadings from implicit learning aptitude to implicit knowledge would be significant, but the path from explicit learning aptitude to implicit knowledge would be non-significant (Hypothesis 5). In contrast, it was expected that the path from explicit learning aptitude to automatized explicit knowledge would be significant, whereas the path from explicit learning aptitude to implicit knowledge would not be significant.

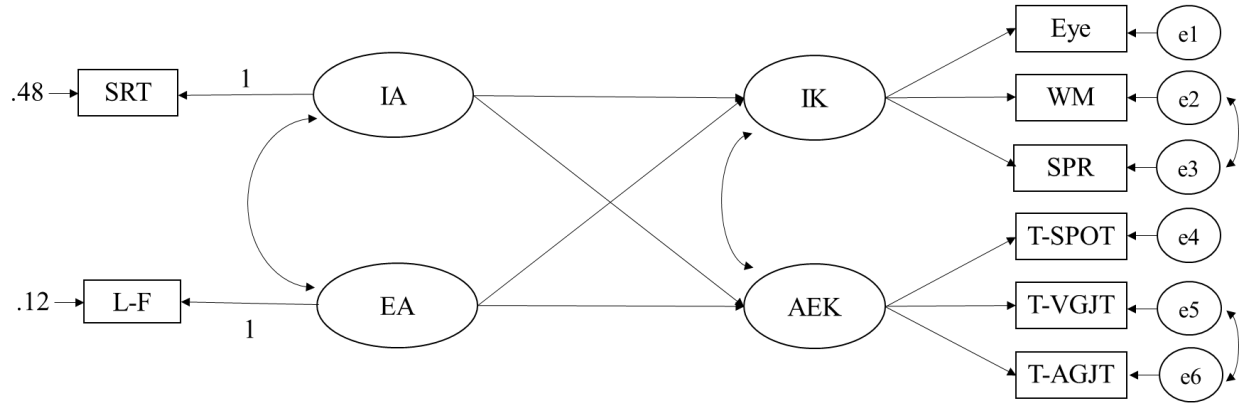


Figure 12. SEM Model 1: Validation Model

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, IA = Implicit Aptitude, EA = Explicit Aptitude, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

5.5.6 The Interface Issue of Explicit and Implicit Learning and Knowledge

After the validation process of the measures for explicit and implicit knowledge, we investigated how the acquisition of two types of knowledge were acquired through SEM analyses. First, the role of cognitive aptitudes for explicit and implicit learning, as well as

[30] The similar pattern of results was obtained with indicators only, instead of modeling the latent factors as predictors.

phonological short-term memory, were examined as predictors for the acquisition of explicit and implicit knowledge. Second, the relationship between the two latent linguistic knowledge was examined.

The two-factor model from Figure 3 was expected to be identified, and three additional aptitude components were added as predictors in SEM analysis (see Figure 13). The direct effects are shown as single-headed arrows, and correlations are shown as double-headed ones. In order to assess the role of individual difference measures, the three latent factors of cognitive aptitudes were posited: implicit learning aptitude, explicit learning aptitude, and memory. As in the SEM model 1, the measurement errors were fixated on the indicators based on the reliability indices of the tests, respectively. The measurement errors were calculated by subtracting the reliability coefficients from 1: the SRT task (*ME* = .48), the letter-span task (*ME* = .08) and the LLAMA F (*ME* = .12).

For the acquisition of explicit knowledge, explicit learning aptitude should play a significant role. The SRT task was expected to predict the acquisition of implicit knowledge. Phonological short-term memory measured with the letter span task would be expected to contribute to the acquisition of both explicit and implicit knowledge because it serves as a basis for explicit and implicit inductive learning. In sum, the following four path loadings were predicted to be significantly different from zero: the one from LLAMA F to AEK, the one from SRT to IK, the one from the letter span to AEK, and the one from the letter span task to IK.

Furthermore, the current study aimed to explore the relationship between explicit and implicit knowledge, i.e., the interface issue. Two models were constructed to empirically test the interface issue: SEM 2 (AEK to IK) and SEM 3 (No interface between AEK and IK). In the SEM 2 model, a path from AEK to IK factor was drawn (Figure 14). This model examined

whether the acquisition of automatized explicit knowledge contributes to the acquisition of implicit knowledge. As a competing model, the SEM model 3 stipulated no relationship between IK and AEK. This model is driven by Krashen's claim that explicit knowledge plays marginal role in the development of implicit knowledge (Krashen, 1981). As in the CFA and MTMM analysis, the maximum likelihood method was used to estimate model parameters. Model fit was evaluated by the multiple indices. As in the other analyses, the correlated errors were added to the model and kept only if they were significant.

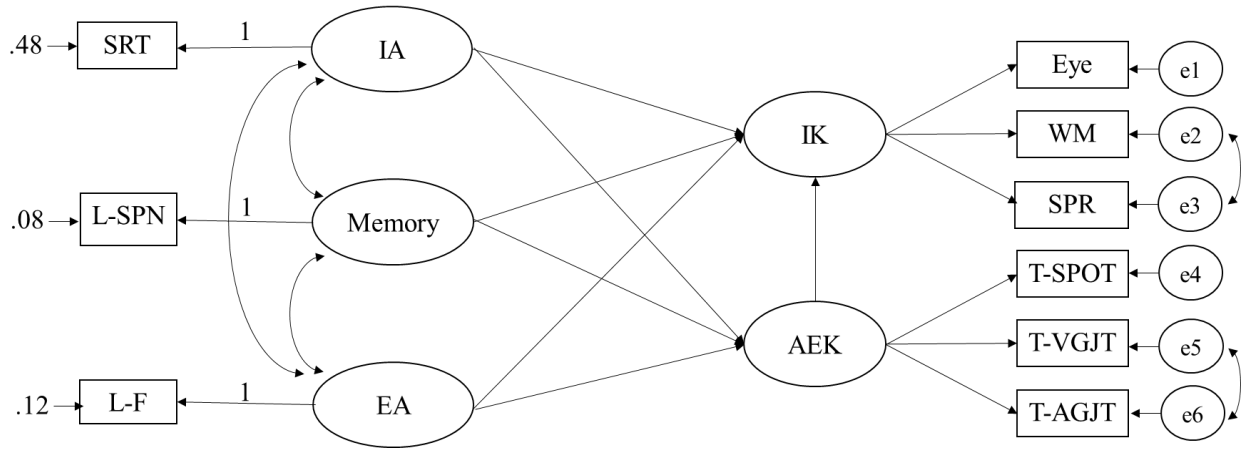


Figure 13. SEM Model 2: AEK to IK model

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

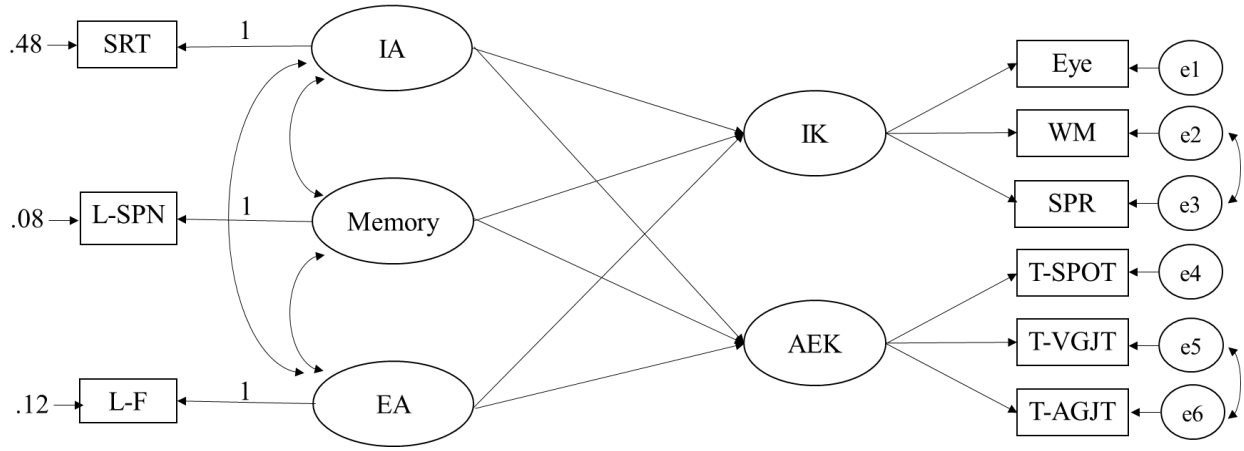


Figure 14. SEM Model 3: No interface model

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

Chapter 6: Results

6.1 Descriptive Statistics for Language Tests

6.1.1 Visual-World Task. Results of the visual-world paradigm are presented for each target structure. In order to demonstrate that the task assessed the targeted linguistic knowledge as designed, NSs' results are presented first, followed by the results of L2 speakers.

6.1.1.1 Transitive/Intransitive. Dependent measures were computed by a proportion of looks to the action doer during the region of interest. The region of interest was from the onset of the case markers (*ga* or *o*) to the onset of NP2. In the region of interest, the proportion of looks to the person was defined as *P* (person) and looks to the contrast object was *P* (contrast). The ratio of *P* (person) to the sum of *P* (person) and *P* (contrast) was computed, and this ranged from 0 (exclusive looks to the contrast), to 0.5 (equal proportion of looks to the contrast and the person), to 1 (exclusive looks to the person). This was compared between the transitive trials and the intransitive trials during the region of interest.

The NSs' proportions of fixations to the person in the transitive and intransitive trials are illustrated in Figure 15. An equal proportion of looks was observed between the transitive and intransitive trials until 400 ms. This is indicated by the overlapped standard error bars between the two trial types. The data-driven (post-hoc) onset of fixations was thus set at 500 ms, and the end of the critical region was set at 1668 ms, corresponding to the offset of the particle *wa.* The paired *t*-test was then performed during this post-hoc critical region (500 – 1668 ms) to assess whether the proportion of looks in the transitive trials was significantly higher than in the intransitive trials. Results showed that the proportion of looks in the transitive trials (M = 63.07%, SD = 9.21%) was significantly greater than in the intransitive trials with a large effect size (M = 48.73%, SD = 6.71%), $t(19) = 6.319$, $p < .001$, $d = 1.768$.

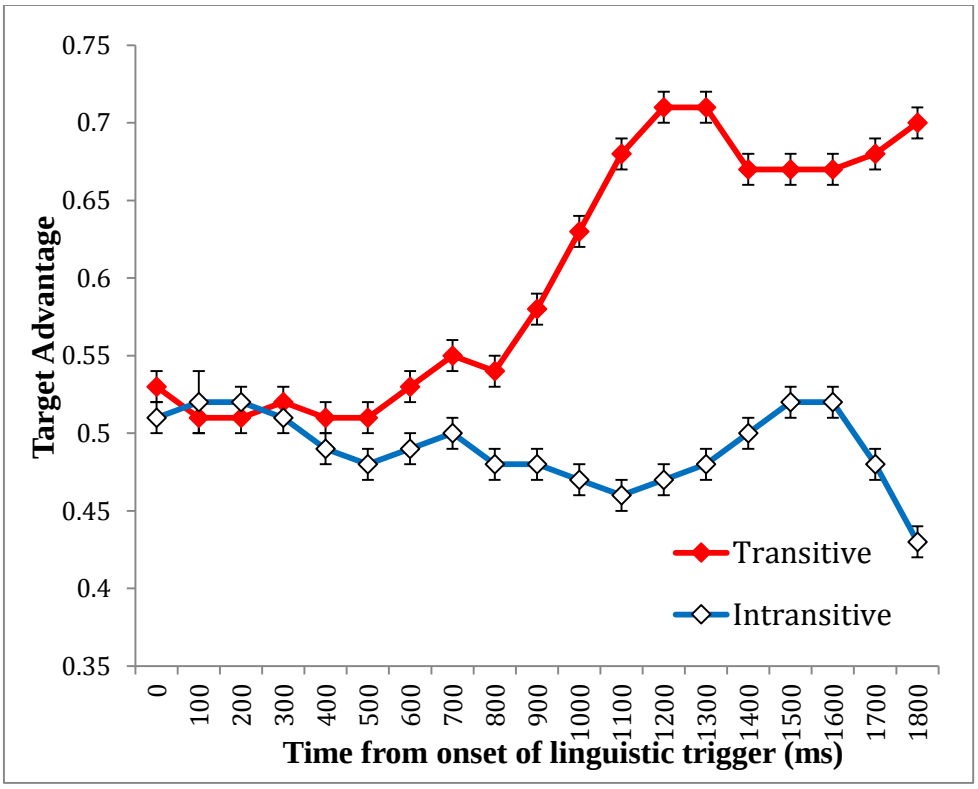


Figure 15. Time-Course of Fixations to Target in Transitive Trials and Intransitive Trials: Native Speakers (n =20)

Figure 16 illustrates the proportion of fixations for the L2 speakers. The looks in the transitive trials became greater than in the intransitive trials at 700 ms, which was slightly slower than for the NSs. Although the proportion of looks was consistently higher for the transitive than for the intransitive, the difference was much smaller in the L2 speakers than in the NSs. That means L2 speakers were less sensitive to the transitivity of verbs than NSs. A paired *t*-test was conducted on the proportion of fixations during the critical region (i.e., 500-1668 ms). There was a significant difference between the transitive trials (M = 52.80%, SD = 7.65%) and the intransitive trials (M = 49.99%, SD = 6.56%) with a medium effect size, $t(99) = 2.797$, $p = .006$,

d = 0.395. In order to compute the sensitivity index to transitivity, the ratio of looks in the intransitive trials was subtracted from the ones in the transitive trials.

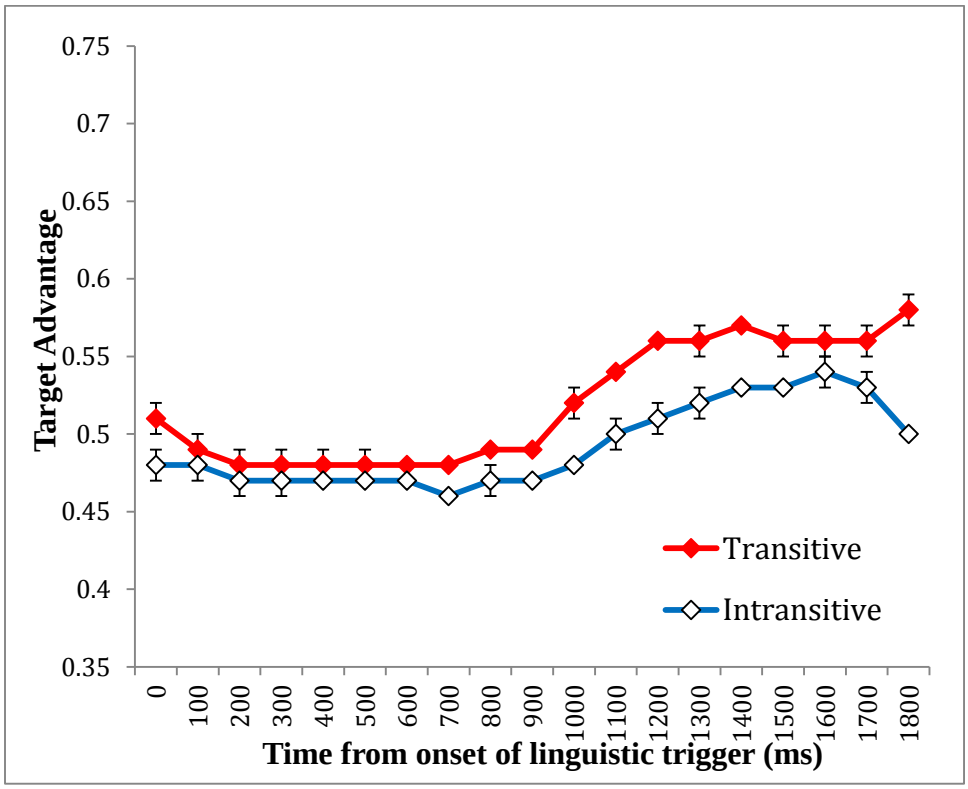


Figure 16. Time-Course of Fixations to Target in Transitive Trials and Intransitive Trials: L2 Speakers (n =100)

6.1.1.2 Classifiers. Dependent measures were computed by a proportion of looks to the target noun during the region of interest. The region of interest was from the onset of the classifier to the onset of the target noun. In the region of interest, the proportion of looks to the target was defined as *P* (target), and that to the competitor noun defined as *P* (competitor). The ratio then was computed between *P* (target) and the sum of *P* (target) and *P* (competitor), and this ranged from 0 (exclusive looks to the competitor), 0.5 (equal proportion of looks to the target and the competitor), to 1 (exclusive looks to the target).

The NSs' proportions of fixations to the target noun are illustrated for the classifier-match trials and the classifier-mismatch trials in Figure 17. The fixation proportions were almost identical at 0 ms between the classifier-match trials and the classifier-mismatch trails. NSs started to look more at the target noun in the classifier-match trials than in the classifier-mismatch trials immediately from 100 ms. The post-hoc onset of fixations was thus set at 100 ms for the classifier construction. The end of the critical region was 1074 ms, which was the onset of the target nouns. The paired *t*-test was then performed during this critical region (100 - 1074 ms) to assess whether the proportion of looks was higher in the classifier-match trials than in the classifier-mismatch trials. Results showed that the proportion of looks in the classifier-match trials (M = 66.41%, SD = 9.95%) was significantly greater than in the classifier-mismatch trials with a large effect size (M = 38.88%, SD = 12.00%), $t(19) = 7.757$, $p < .001$, $d = 2.498$.

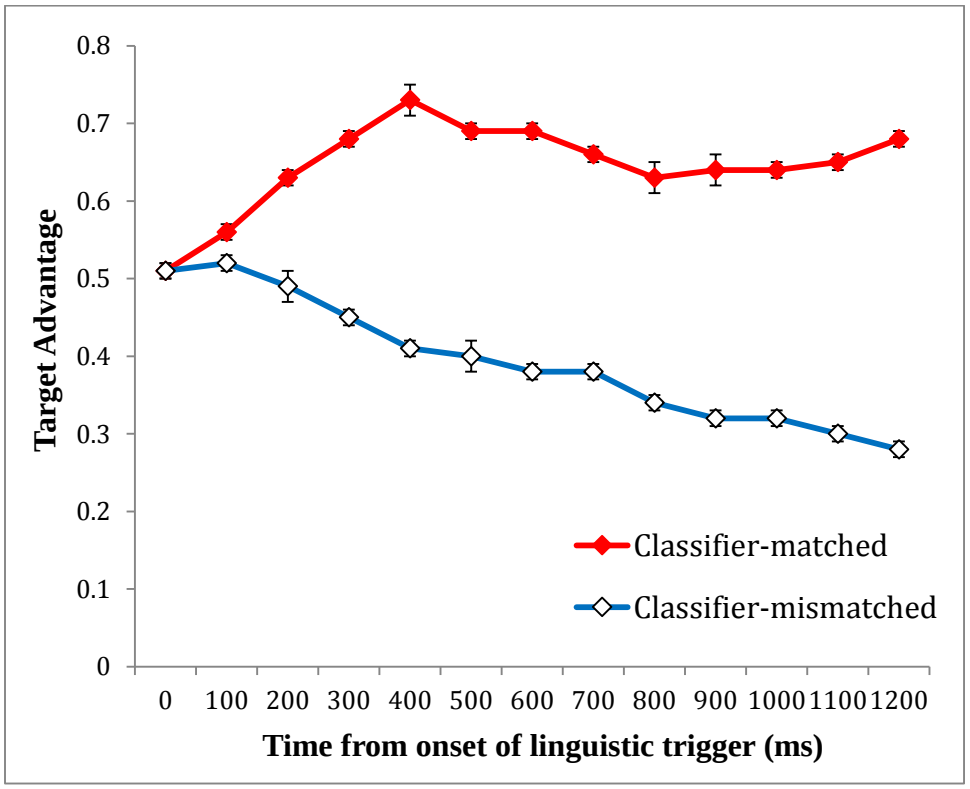


Figure 17. Time-Course of Fixations to Target in Classifier-matched Trials and Classifier-mismatched trials: Native Speakers (n =20)

Figure 18 illustrates the proportion of fixations for the L2 speakers. The looks to the target became greater in the classifier-matched trials than in the classifier-mismatched trials at approximately 300 ms, with the difference appearing to be smaller in the L2 speakers. A paired *t*-test was conducted on the proportion of fixations during the critical region (i.e., 100-1074 ms). There was a significant difference between the proportion of looks to the target in the target trials (M = 62.02%, SD = 9.59%) and in the baseline trials (M = 45.51%, SD = 12.26%) with a large effect size, $t(99) = 11.384$, $p < .001$, $d = 1.497$. The difference score between the ratio of looks in the target trials and that in the baseline trials was used as an index for the acquisition of the classifier. In order to compute the acquisition index for the classifier structure, the ratio of looks in the classifier-mismatched trials was subtracted from the ones in the classifier-matched trials.

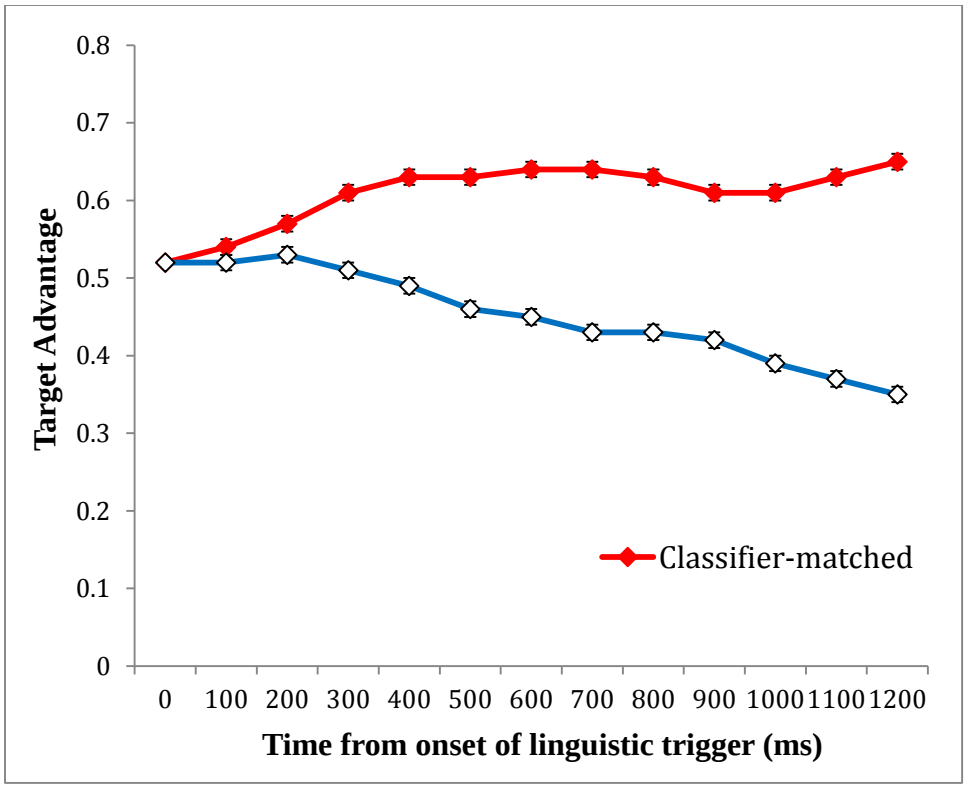


Figure 18. Time-Course of Fixations to Target in Classifier-matched Trials and Classifier-mismatched trials: L2 Speakers (n =100)

6.1.1.3 Ni/De. The dependent variable was the proportion of fixations on Person A (performer) during the region of interest. The region of interest was from the onset of the particle, *ni* or *de*, to the onset of the VP. In the region of interest, the proportion of looks to the person was defined as *P* (person A) and that to a referent object in the *ni* condition as *P* (person B). The ratio of *P* (person A) to the sum of *P* (person A) and *P* (person B) was computed, and this ranged from 0 (exclusive looks to Person B), to 0.5 (equal proportion of looks to Person A and Person B), to 1 (exclusive looks to Person A).

The NSs' proportions of fixations to Person A (i.e., performer) in the *de* and the *ni* trials are illustrated in Figure 19. NSs looked at the target equally at 0 ms between *de* and *ni* trials. The NSs' eye-movements started to be directed more at the target from 100 ms. The post-hoc onset

of fixations was thus set at 100 ms for the *ni*/*de* construction, and the end of critical region was at the onset of the VP at 1074 ms. The paired *t*-test was then performed during this critical region (100 – 1074 ms). Results showed that the proportion of looks in the *de* trials (M = 58.55%, SD = 11.83%) was significantly greater than in the *ni* trials with a large effect size (M = 41.65%, SD = 7.48%), $t(19) = 5.119$, $p < .001$, $d = 1.717$.

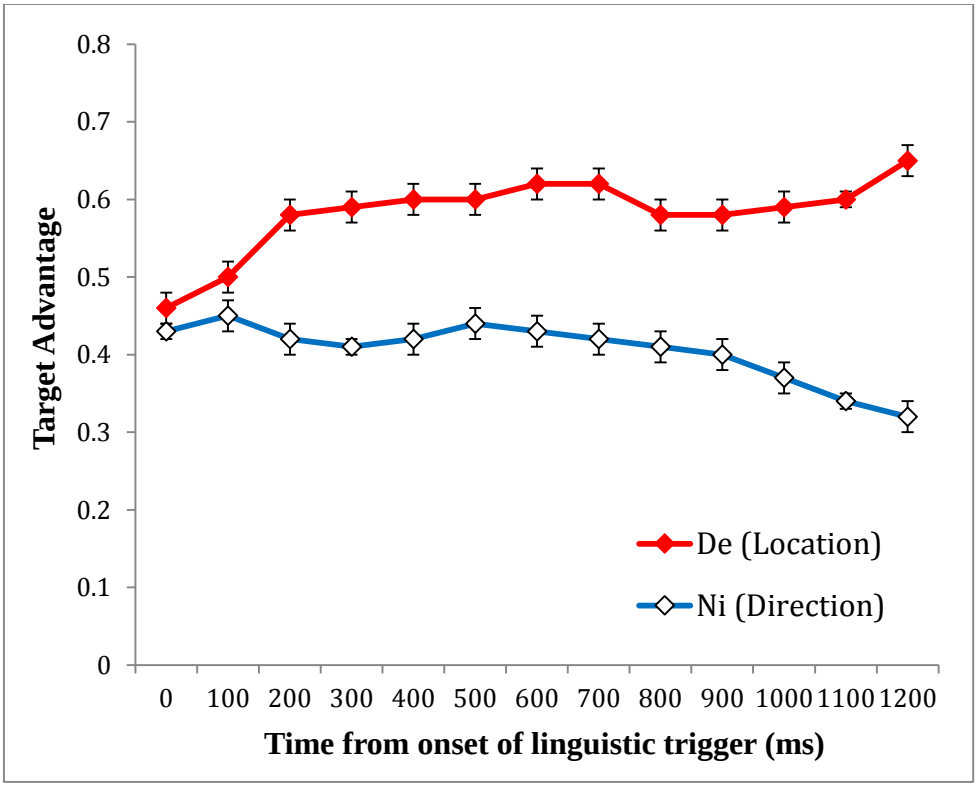


Figure 19. Time-Course of Fixations to Target in *De* Trials and *Ni* Trials: Native Speakers (n =20)

Figure 20 illustrates the proportion of fixations for the L2 speakers. L2 speakers started to look at Person A more in the *de* trials than the *ni* trials from 500 ms. The paired *t*-test was conducted on the proportion of fixations during the critical region (i.e., 100-1074 ms). There was

a significant difference between the *de* trials (M = 54.09%, SD = 13.21%) and the *ni* trials (M = 49.53%, SD = 11.91%) with a medium effect size, $t(99) = 2.498$, $p = .014$, $d = 0.363$.

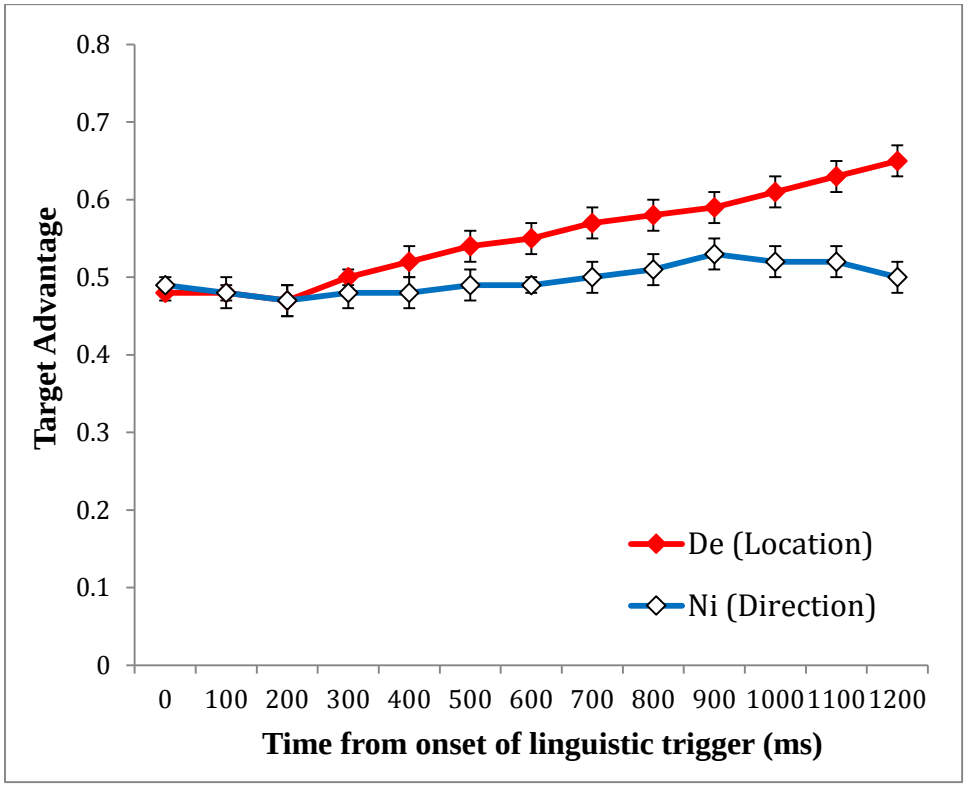


Figure 20. Time-Course of Fixations to Target in *De* Trials and *Ni* Trials: L2 Speakers (n =100)

6.1.1.4 Tameni/Youni. Dependent variables were the proportion of fixations to person B (the person who is mentioned in the main clause followed by *youni*). The region of interest was from the onset of *tameni/youni* to the offset of the object in the main clause (e.g., *jaaji o*), which corresponded to the onset of Person B in the *youni* sentences. In the region of interest, the proportion of looks to Person B was defined as *P* (person B) and that to Person A as *P* (person A). The ratio of *P* (person B) to the sum of *P* (person B) and *P* (person A) was compared between the *tameni* sentences and the *youni* sentences, and this ranged from 0 (exclusive looks to person A), to 0.5 (equal proportion of looks to person A and person B), to 1 (exclusive looks to person B).

The NSs' proportions of fixations to the target (i.e., person B) in the *tameni* and the *youni* trials are illustrated for NSs in Figure 21. NSs looked at the target equally in the *tameni* and *youni* trials approximately until 400 ms, and then at 500 ms, they started to be look at the target consistently more in the *youni* trials than in the *tameni* trials. The post-hoc onset of fixations was thus set at 500 ms, and the end of the critical region was 1428 ms, which was the offset of the object in the main clause. The paired *t*-test was then performed during this critical region (500-1428 ms). Results showed that the proportion of looks to person B in the *youni* trials (M = 57.18%, SD = 12.74%) was significantly greater than in the *tameni* trials with a large effect size (M = 45.88%, SD = 13.61%), $t(19) = 2.397$, $p = .027$, $d = 0.858$.

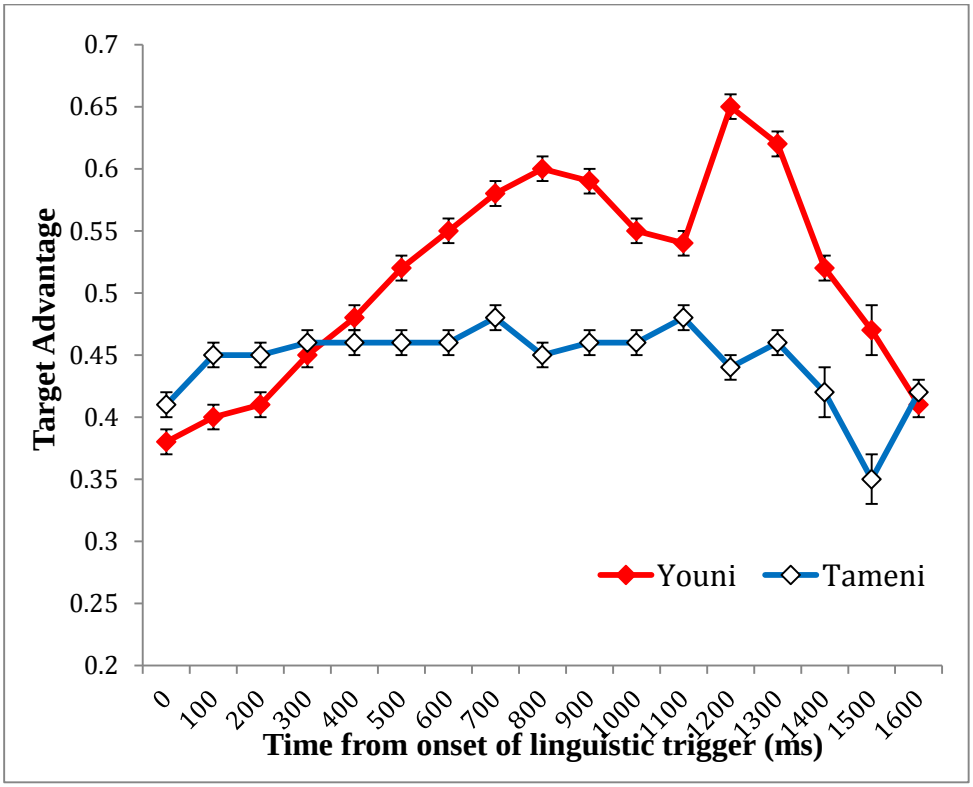


Figure 21. Time-Course of Fixations to Target in *Youni* Trials and *Tameni* Trials: Native Speakers (n =20)

Figure 22 illustrates the proportion of fixations for the L2 speakers. Unlike NSs, L2 speakers did not show any sensitivity to the distinction between *tameni* and *youni* at all until 1000 ms, and the looks in the *tameni* trials became greater than in the *youni* trials. This pattern was opposite to the NSs' pattern; the paired *t*-test for the critical region (i.e., 500-1428 ms), however, showed no significant difference between the *youni* trials (M = 47.50%, SD = 10.68%) and the *tameni* trials (M = 48.86%, SD = 11.05%), $t(99) = 0.877$, $p = .383$, $d = 0.126$.

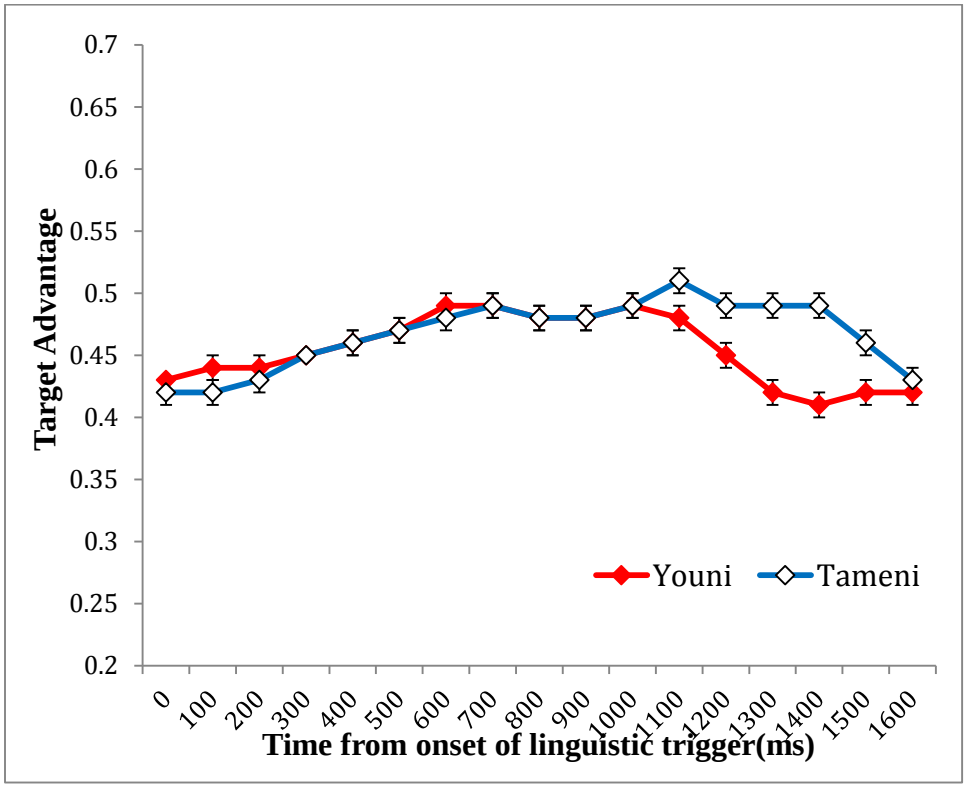


Figure 22. Time-Course of Fixations to Target in *Youni* Trials and *Tameni* Trials: L2 Speakers (n =100)

6.1.2 Word-Monitoring Task. In order to show the task was designed well and appropriately performed by the NSs, paired-sample t-tests were performed to compare the difference in RTs between grammatical and ungrammatical items (see Table 12). A significant difference was detected with a large effect size for all the structures except for *tameni/youni*.

Although NSs showed sensitivity to the difference between *tameni* and *youni* in the visual-world task, they were not sensitive to the distinction in the word-monitoring task. For the L2 group, a paired-samples t-test showed a significant difference with a small effect size for classifiers and *ni/de*; no significant difference was found in the other two structures.

Table 12. *Descriptive Statistics and Results of T-tests on Mean RTs for Grammatical and Ungrammatical Items by Target Structures across Two Groups*

		Tran/Intran	Classifier	Ni/De	Tam/You	All[1]
NSs (n = 24)						
Gram	Mean (SD)	393(78)	374(86)	430(65)	452(452)	399(66)
	Range	248-539	229-566	283-556	306-573	253-513
Ungram	Mean (SD)	516(97)	467(68)	529(107)	448(448)	504(83)
	Range	320-746	314-634	346-815	352-681	326-707
	Difference	123	94	100	-5	105
	t	8.546	7.614	6.082	-.369	13.099
	p	.000	.000	.000	.716	.000
	Cohen's *d*	1.361	1.163	1.019	-0.061	1.283
L2 Speakers (n =100)						
Gram	Mean (SD)	507(140)	477(132)	496(138)	483(483)	493(124)
	Range	259-958	252-944	246-1024	253-930	264-861
Ungram	Mean (SD)	510(144)	518(132)	518(150)	477(477)	515(129)
	Range	300-1149	301-1131	309-1199	289-1440	325-1023
	Difference	4	41	22	-5	22
	t	0.341	4.428	2.509	-.498	4.056
	p	.734	.000	.014	.619	.000
	Cohen's *d*	0.025	0.310	0.151	-0.035	0.174

Note 1. Mean RTs were computed for all the structures except for *tameni/youni* because the mean RTs for grammatical and ungrammatical sentences were not statistically significant for *tameni/youni* even in NSs' data.

6.1.3 Self-Paced Reading Task.

6.1.3.1 Transitive/Intransitive. Mean RTs for grammatical and ungrammatical sentences with transitive/intransitive for each critical region are presented in Table 13. Paired t-tests revealed a significant difference only at Region 2 with a small effect size in the NS group. No significant differences were found for the L2 speakers at any regions.

Table 13. *Descriptive Statistics and Results of T-tests on Mean RTs for Grammatical and Ungrammatical Sentences with Transitive/Intransitive Structures Measured at Four Positions across Two Groups*

	Tran/Intran			
Region	0	1	2	3
NSs (n = 16)				
Gram	276(87)	272(87)	283(75)	300(85)
Ungram	288(75)	279(82)	319(96)	312(73)
Difference	12	7	36	12
t	1.132	1.148	2.637	0.816
p	.275	.269	.019	.427
Cohen's *d*	0.140	0.080	0.391	0.146
L2 speakers (n = 100)				
Gram	519(143)	571(171)	561(148)	539(129)
Ungram	531(141)	564(169)	574(155)	536(131)
Difference	12	-7	13	-4
t	1.150	-0.644	0.938	-0.402
p	.253	.521	.351	.689
Cohen's *d*	0.084	-0.043	0.084	-0.030

6.1.3.2 Classifiers. Mean RTs for grammatical and ungrammatical sentences with classifiers for each critical region are presented in Table 14. No significant difference was found at Region 0 for the NSs and L2 speakers. Paired t-tests revealed a marginally significant difference at Region 2 and a significant difference at Region 3 in the NS group. Effect sizes for both regions were small. Similarly, significant differences were detected at Region 2 and Region 3 with small effect sizes for the L2 groups.

Table 14. *Descriptive Statistics and Results of T-tests on Mean RTs for Grammatical and Ungrammatical Sentences with Classifiers Measured at Four Positions across Two Groups*

	Classifiers			
Region	0	1	2	3
NSs (n = 24)				
Gram	319(102)	310(84)	305(80)	300(75)
Ungram	306(89)	299(93)	328(112)	326(93)

Difference	-13	-11	23	26
t	-1.206	-0.978	1.874	3.383
p	.240	.338	.074	.003
Cohen's *d*	-0.130	-0.123	0.203	0.265
L2 speakers (n = 100)				
Gram	505(137)	539(152)	520(118)	496(110)
Ungram	494(134)	557(179)	563(145)	540(134)
Difference	-12	18	43	44
t	-1.427	1.673	4.301	4.782
p	.157	.097	.000	.000
Cohen's *d*	-0.085	0.108	0.319	0.349

6.1.3.3 Ni/De. Mean RTs for grammatical and ungrammatical sentences with *ni/de* for each critical region are presented in Table 15. No significant difference was found at Region 0 for the NSs and L2 speakers. Paired t-tests revealed a significant difference at Region 2 and Region 3 with a small effect size in the NS group. For the L2 group, there were marginally significant differences at Region 1 and Region 2 with small effect sizes for the L2 group.

Table 15. *Descriptive Statistics and Results of T-tests on Mean RTs for Grammatical and Ungrammatical Sentences with Ni/De Measured at Four Positions across Two Groups*

	Ni/De			
Region	0	1	2	3
NSs (n = 24)				
Gram	299(94)	312(97)	312(90)	294(68)
Ungram	313(105)	327(99)	361(123)	325(77)
Difference	14	15	49	31
t	1.388	1.561	3.855	4.485
p	.178	.132	.001	.000
Cohen's *d*	0.138	0.151	0.395	0.409
L2 speakers (n = 100)				
Gram	500(141)	557(156)	508(120)	476(108)
Ungram	503(135)	580(170)	527(130)	480(113)
Difference	2	22	19	4
t	0.288	1.901	1.668	0.450
p	.774	.060	.098	.653

Cohen's *d*	0.018	0.136	0.149	0.033

6.1.3.4 Tameni/Youni. Mean RTs for grammatical and ungrammatical sentences with *tameni/youni* for each critical region are presented in Table 16. No significant difference was found at Region 0 for the NSs and L2 speakers. As found in the word-monitoring task, paired *t*-tests did not show any significant difference at any of the critical regions; the NS group did not show sensitivity to the distinction. No significant differences were detected at any regions for the L2 group, either.

Table 16. *Descriptive Statistics and Results of T-tests on Mean RTs for Grammatical and Ungrammatical Sentences with Tameni/Youni Measured at Four Positions across Two Groups*

	Tameni/Youni			
Region	0	1	2	3
NSs (n = 24)				
Gram	305(110)	307(88)	315(73)	307(75)
Ungram	295(87)	321(98)	307(79)	301(78)
Difference	-10	13	-8	-6
t	-1.01	1.32	-0.81	-0.74
p	.323	.200	.426	.466
Cohen's *d*	-0.09	0.14	-0.11	-0.08
L2 speakers (n = 100)				
Gram	563(165)	572(153)	494(110)	492(117)
Ungram	562(164)	566(155)	489(104)	480(106)
Difference	-1	-7	-6	-12
t	-0.108	-0.488	-0.640	-1.544
p	.914	.627	.523	.126
Cohen's *d*	-0.006	-0.042	-0.053	-0.106

6.1.4 Auditory GJT. Tables 17 and 18 present the total percentage scores of the auditory GJT (raw and timed scores), as well as the scores collapsed by grammatical and ungrammatical items. For the raw scores (before applying the cut-off for RT), NSs' performance was above 90% across all the structures except for *tameni/youni*. The NSs accepted grammatical sentences

accurately, but they rarely rejected ungrammatical items for *tameni/youni.* For the timed version of the test, NSs' total scores were slightly lower but all above 90% except for the *tameni/youni*. These confirmed that the auditory GJT was designed appropriately. The L2 speakers' raw mean scores were slightly above 50% across the structures, and the scores on the timed version ranged from 40% to 50%.

Table 17. *Descriptive Statistics for the Auditory GJT (Raw Score)*

		Tran/Intran	Classifier	Ni/De	Tam/You
NSs (n = 31)					
Gram	Mean (SD)	96.37(6.61)	97.06(5.54)	98.79(3.76)	97.98(5.68)
	Range	75-100	85.71-100	87.5-100	75-100
Ungram	Mean (SD)	92.74(11.54)	87.38(16.2)	86.69(21.87)	2.82(5.31)
	Range	50-100	37.5-100	0-100	0-12.5
Total	Mean (SD)	94.56(6.8)	92.04(8.15)	92.74(10.96)	50.4(3.59)
	Range	68.75-100	66.67-100	50-100	43.75-56.25
L2 Speakers (n = 100)					
Gram	Mean (SD)	87.38(13.35)	84.91(15.45)	84.63(16.93)	82.25(19.23)
	Range	50-100	42.86-100	37.5-100	12.5-100
Ungram	Mean (SD)	21.13(21.66)	32.82(29.12)	23.25(21.69)	21.25(20.91)
	Range	0-100	0-100	0-87.5	0-87.5
Total	Mean (SD)	54.25(10.2)	58.53(15.24)	53.94(11.12)	51.75(11.13)
	Range	37.5-87.5	33.33-100	31.25-87.5	25-81.25

Table 18. *Descriptive Statistics for the Auditory GJT (with cut-off)*

		Tran/Intran	Classifier	Ni/De	Tam/You
NSs (n = 31)					
Gram	Mean (SD)	96.37(6.61)	96.6(5.87)	98.79(3.76)	97.98(5.68)
	Range	75-100	85.71-100	87.5-100	75-100
Ungram	Mean (SD)	89.11(15.05)	83.99(18.04)	84.27(21.64)	2.82(5.31)
	Range	37.5-100	37.5-100	0-100	0-12.5
Total	Mean (SD)	92.74(8.7)	90.11(9.09)	91.53(10.77)	50.4(3.59)
	Range	62.5-100	66.67-100	50-100	43.75-56.25
L2 Speakers (n = 100)					
Gram	Mean (SD)	75.38(20.91)	75.23(21.34)	74.38(21.5)	72.88(23.91)
	Range	25-100	12.5-100	12.5-100	0-100
Ungram	Mean (SD)	6.88(14.69)	16.95(24.76)	11.75(18.45)	4.13(7.55)
	Range	0-75	0-100	0-87.5	0-25
Total	Mean (SD)	41.13(12.82)	45.8(16.75)	43.06(13.79)	38.5(11.71)

	Range	12.5-81.25	6.67-93.33	12.5-81.25	0-62.5

6.1.5 Visual GJT. Tables 19 and 20 present the total percentage scores of the visual GJT (raw and timed scores), as well as the scores collapsed by grammatical and ungrammatical items. As in the auditory GJT, NSs' raw scores were above 90% across all the structures except for *tameni/youni*. The NSs accepted grammatical sentences accurately, but they also failed to reject almost all the ungrammatical items for *tameni/youni*. For the timed version of the test, NSs' total scores were slightly lower but all above 90% except for the *tameni/youni*. These confirmed that the visual GJT was designed appropriately. The raw mean scores by L2 speakers were just above 50% for the transitive and the *ni/de* distinction, and the score for the classifier was above 60%. The mean scores for the time version ranged 20% to 40%.

Table 19. *Descriptive Statistics for the Visual GJT (Raw Score)*

		Tran/Intran	Classifier	Ni/De	Tam/You
NSs (n = 31)					
Gram	Mean (SD)	88.36(15.82)	90.73(16.76)	98.39(4.26)	97.98(4.67)
	Range	42.86-100	25-100	87.5-100	87.5-100
Ungram	Mean (SD)	94.82(6.66)	90.32(13.58)	85.48(20.94)	2.42(5.02)
	Range	85.71-100	50-100	25-100	0-12.5
Total	Mean (SD)	91.61(8.25)	90.52(8.82)	91.94(10.48)	50.2(3.42)
	Range	73.33-100	62.5-100	62.5-100	43.75-56.25
L2 Speakers (n = 100)					
Gram	Mean (SD)	35.43(28.34)	62.75(26.71)	71.75(20.38)	66(26.65)
	Range	0-100	0-100	12.5-100	0-100
Ungram	Mean (SD)	67.96(21.62)	62.75(28.43)	36.13(26.47)	32.63(26.94)
	Range	14.29-100	0-100	0-100	0-100
Total	Mean (SD)	51.87(12.77)	62.75(17.32)	53.94(13.78)	49.31(11.68)
	Range	26.67-93.33	25-100	25-93.75	18.75-75

Table 20. *Descriptive Statistics for the Visual GJT (with cut-off)*

		Tran/Intran	Classifier	Ni/De	Tam/You
NSs (n = 31)					
Gram	Mean (SD)	87.1(17.14)	90.32(17.29)	97.98(4.67)	97.58(5.02)
	Range	42.86-100	25-100	87.5-100	87.5-100

Ungram	Mean (SD)	94.35(7.72)	89.92(13.85)	84.27(20.91)	2.42(5.02)
	Range	71.43-100	50-100	25-100	0-12.5
Total	Mean (SD)	90.75(8.89)	90.12(8.95)	91.13(10.55)	50(3.61)
	Range	73.33-100	62.5-100	62.5-100	43.75-56.25
L2 Speakers (n = 100)					
Gram	Mean (SD)	15.13(21.42)	32.88(28.07)	41.63(27.3)	39.38(29.11)
	Range	0-100	0-100	0-100	0-100
Ungram	Mean (SD)	35.98(24.72)	38.38(30.11)	19.88(22.9)	5.25(7.98)
	Range	0-100	0-100	0-87.5	0-25
Total	Mean (SD)	25.6(16.02)	35.63(21.4)	30.75(18)	22.31(13.68)
	Range	0-73.33	0-100	0-87.5	0-50

6.1.6 SPOT. Table 21 and 22 present the percentage scores for the SPOT. NSs' raw scores were near 100% across all the structures except for *tameni/youni*. The NSs showed sensitivity to the *tameni/youni* distinction in the SPOT, which contrasts with the results in all the tasks except for the visual-world task. NSs do seem to distinguish *tameni* and *youni* in the tasks in which ungrammatical sentences were not included, but it must have been too subtle to make grammatical judgments or to show sensitivity to ungrammatical sentences.

For the timed version of the test, the NSs' total scores were slightly lower but all above 88% except for the *tameni/youni*. These confirmed the SPOT appropriately assessed the linguistic knowledge of the target structures. The L2 speakers' raw mean scores were higher than the two GJTs; the raw mean scores ranged from 66% to 86%. The mean scores on the time version were much lower, ranging from 20% to 40%.

Table 21. *Descriptive Statistics for the SPOT (Raw Score)*

	Tran/Intran	Classifier	Ni/De	Tam/You
NSs (n = 31)				
Mean (SD)	98.79(2.51)	99.8(1.12)	100(0)	82.86(12.18)
Range	93.75-100	93.75-100	100-100	50-100
L2 Speakers (n = 100)				
Mean (SD)	66.66(17.75)	78.38(19.37)	86.19(13.06)	74.11(17.03)
Range	18.75-100	25-100	50-100	37.5-100

Table 22. *Descriptive Statistics for the SPOT (with cut-off)*

	Tran/Intran	Classifier	Ni/De	Tam/You
NSs (n = 31)				
Mean (SD)	88.51(19.17)	92.74(15.57)	90.52(20.34)	74.6(18.04)
Range	12.5-100	37.5-100	6.25-100	12.5-93.75
L2 Speakers (n = 99)				
Mean (SD)	23.11(22.67)	26.26(24.34)	32.01(28.25)	39.27(24.95)
Range	0-87.5	0-93.75	0-100	0-100

6.2 Descriptive Statistics for Cognitive Aptitude Tests

6.2.1 LLAMA F. The scores of LLAMA F were based on the sum of correct responses. Five participants were excluded from the analyses that involve LLAMA scores due to experimental errors or the fact that participants did not follow the instructions. The mean was 23.18 (*SD* = 4.19). According to the Kolmogorov-Smirnov tests, the distribution was not normal (Kolmogorov-Smirnov = .152, $p < .001$).

6.2.2 SRT Task. Before computing the Serial Reaction Time (SRT) scores for each participant, error responses were discarded (2% of trials), and outliers that were three SDs from the mean for each participant were also discarded (1.6% of trials).

Mean RTs for the probable condition (85%) and for the improbable condition (15%) across blocks are presented in Figure 23. A repeated-measures analysis of variance (ANOVA) was conducted on the RT with block and condition (probable versus non-probable) as within-subject factors. According to Mauchly's test, the assumption of sphericity was violated for block, $\chi^2(35) = 368.427$, $p < .001$, and for block*random, $\chi^2(35) = 132.616$, $p < .001$; therefore, the results below are reported with the Greenhouse-Geisser correction. They show a significant effect of block, F (3.506, 347.077) = 35.537, $p < .001$, $\eta^2 = 0.264$, and condition, F (1, 99) = 48.422, $p < .001$, $\eta^2 = 0.328$. A significant interaction between block and condition was also detected, F (5.965, 590.529) = 44.921, $p < .001$, $\eta^2 = 0.312$. Figure 23 shows that the learning

effect was established at block 3, which concurred with the pattern found in Kaufman et al. (2010) and Suzuki and DeKeyser (in press), which used the same SRT task. The amount of implicit learning was calculated from the third block, in which the effect was established, to the last block. A paired-samples t-test revealed a significant RT difference across the last six blocks between the two conditions, $t(99) = 10.491$, $p < .001$, $d = 0.233$. According to the K-S tests, the distribution of SRT was normal ($p > .05$).

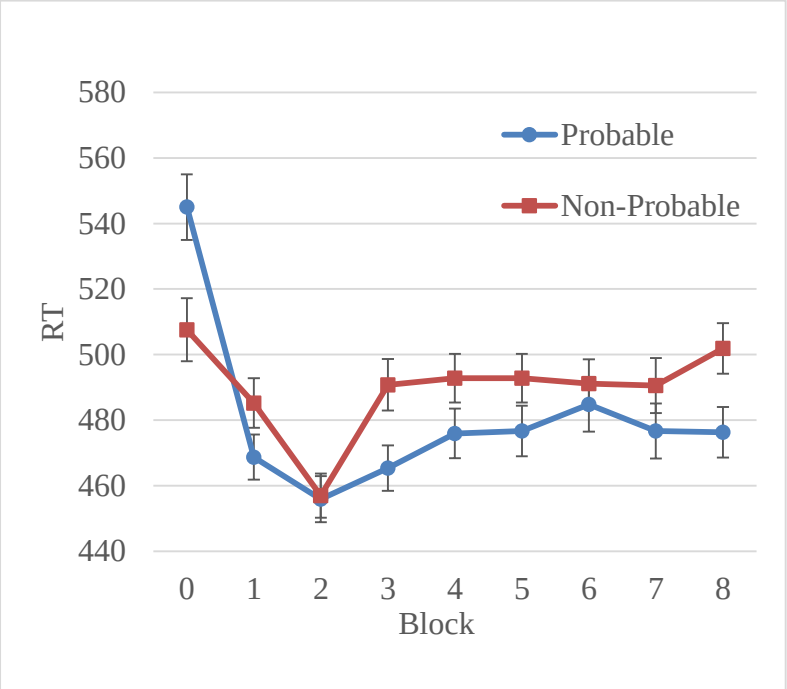


Figure 23. SRT task learning performance in probable and improbable trials

In order to examine whether participants developed explicit knowledge about sequence knowledge during the SRT task, a recognition test was administered immediately after the SRT task. The performance on the test consisting of a subjective component (confidence ratings) and an objective component (RT) was analyzed. First, participants' confidence ratings given to old (i.e., the more probable sequence A) and new sequences (i.e., the less probable sequence B) were compared (Table 23). According to a paired-sample t-test, no significant difference was found between the old and new sequences with a negligible effect size, $t(99) = 0.319$, $p = .751$, $d =$

0.024. Lack of conscious discrimination of old sequences from new sequences suggests that the SRT performance reflected implicit learning with little influence of explicit knowledge.

Second, the RT on the third dot of the same old and new sequences was also compared, after excluding outliers (1.13%) that were beyond 3 SD above or below each participant's mean RT (see Table 23). Faster RT in the old sequences than in the new sequences provides "a direct index of the possible influence of unconsciously applied perceptual-motor programs" (Shanks & Johnstone, 1999, p. 1446). A paired-sample t-test revealed a significant difference between the old and new sequences, $t(99) = 2.043$, $p = .044$, $d = 0.108$. This ensured, in combination with the confidence ratings results, that participants developed implicit knowledge of sequences without explicit knowledge.

Table 23. *Descriptive Statistics for the Recognition Test: Confidence Ratings and RT*

	Sequences	Mean	SD	Min	Max
Confidence Rating	Old	2.28	0.65	1.00	4.00
	New	2.30	0.66	1.00	3.67
RT	Old	501	143	294	1141
	New	516	131	355	1122

Note: Lower confidence ratings on a six-point scale indicated greater confidence in the sequence being old.

6.2.3 Letter-Span Task. The score of the letter-span task was calculated by the total number of letters recalled in their correct positions. The mean score was 89.87 ($SD = 14.61$). According to the K-S tests, the distribution of SRT was normal (K-S = .055, $p > .05$).

6.3 Data Preparation for CFA and SEM Analyses

6.3.1 Data Summary. For the main CFA and SEM analyses, language test scores were combined across all the target structures but *tameni/youni*. Since the visual-world task and the self-paced reading assessed the use of linguistic knowledge over the time course of sentence

processing, several decisions were made to determine the time window for computing the index for linguistic knowledge.

In the visual-world task, the proportions of fixations were time-locked from the time in which native speakers demonstrated using the linguistic trigger reliably (500 ms for transitive and 100 ms for classifiers and *ni/de*) to the end of critical region (1668 ms for transitive and 1074 ms for classifiers and *ni/de*). Since the critical region is long (i.e., around 1000 ms), we decided to select the narrower time region to capture the rapid use of linguistic knowledge. Autocorrelations among the fixation proportions (for the three structures combined) at 100 ms were computed in Table 24. The proportion of looks in the 100 ms time window was internally consistent particularly in the beginning, indicated by the higher correlations, and the correlation with the looks in the later regions became smaller and smaller. This suggests that the looks to the items changed during the critical region as listeners' fixations might not always dwell on the same picture once they looked at it.

There could not be any theoretical decision to determine the duration of the time window in which implicit knowledge is likely to be deployed. It was decided to combine only the first 200 ms, however. Three justifications were offered for this decision. First, implicit knowledge should be deployed very rapidly. It is assumed that setting a longer time window might increase the chance of including the use of explicit knowledge, although use of explicit knowledge is unlikely, as participants' attention is directed to meaning. In the SRT task, explicit knowledge was more likely to be developed when the inter-stimulus interval was set at 250 ms than 0 ms (Destrebecqz & Cleeremans, 2001). Second, reaction-time tasks, particularly the word-monitoring task, are pre-time-locked, in order for the online sensitivity to be revealed. The errors were embedded right before the target word, and error detection should happen almost at the

exact point in time where ungrammatical segments occur in speech. As the first justification delineates, it is important to capture the earliest sensitivity to the target linguistic structures. Third, the fixations during the earlier time regions are most likely to be driven by the linguistic trigger more directly than the later time region, which is partially demonstrated by the fact that the correlations of the fixation proportions during the first 100 ms with those at later time windows became lower and lower. Given the highest correlation between the fixations during 100 and 200 ms (r = .779) and the two accounts above, the index was computed by collapsing the first two 100 ms windows (see Appendix E for further discussion).

Table 24. *Autocorrelations among the Fixation Proportions by 100 ms (L2 speakers, n = 100)*

	1. 100ms	2. 200ms	3. 300ms	4. 400ms	5. 500ms	6. 600ms	7. 700ms	8. 800ms	9. 900ms
1	-	.779**	.538**	.490**	.446**	.294**	.276**	.262**	.225*
2		-	.821**	.685**	.597**	.450**	.402**	.336**	.312**
3			-	.771**	.682**	.542**	.454**	.338**	.345**
4				-	.885**	.682**	.613**	.486**	.434**
5					-	.858**	.749**	.598**	.526**
6						-	.889**	.743**	.689**
7							-	.912**	.815**
8								-	.923**
9									-

For the self-paced reading task, since there were three critical positions, different ways of computing indices were available. For the purpose of the present study, and applying the similar rationale as for the visual-world task, the first two regions were combined to index the earliest sensitivity to target structures. The RT difference between the grammatical and ungrammatical sentence at Region 1 (i.e., where the ungrammatical feature occurs) was included to capture the

earliest sensitivity.[31] This was considered to be an effective decision because the first critical word was also at the same location as the target word in the word-monitoring task, which makes the index comparable between the tasks. In order to capture the spillover, RT differences at Region 2 were included, but those at Region 3 were not.[32] In a previous self-paced reading task study, Jiang et al. (2011) combined the second and the third regions in order to capture the individual differences in the location in which L2 speakers show the difference in RTs for grammatical and ungrammatical sentences. The main focus of Jiang's study was to examine whether L2 speakers were ever sensitive to the ungrammatical errors including the later region (e.g., Region 3). Note that NSs did not show significant RT differences at Region 1 in any of the target structures, but L2 speakers seemed to show the larger differences at Region 1 for the classifiers and *ni*/*de*. This lent further support for including the earliest region to compute the index for L2 speakers. NSs did not show significant RT differences, probably because they read words and pressed the button very quickly (i.e., around 300 ms), and the slowdown in reading spilled over to the subsequent regions. Since reading times were recorded by the button presses, the artifact of this method might lead to less-reliable estimations of reading times.

Descriptive statistics for all the measures are presented in Table 25. Reliability indices were above .80, except for the three tests: the visual-world task, the timed AGJT, and the SRT

[31] The use of earlier time points might be supported by a recent ERP study. Batterink and Neville (2013) claim that unconscious detection, which is the hallmark of morphosyntactic processing without awareness, was associated with the earlier neural responses (100-400 ms). Unconscious detection lacked the later neural response during 900-1200 ms, which was accompanied by the conscious identification of grammatical violations. The findings suggest that the earlier responses to the grammatical violation seems to be associated with implicit processing rather than explicit processing. However, another recent study casts doubt on Batterink and Neville's findings on the ground that an early left anterior negativity (LAN, usually manifested during 300-500 ms) might be an artifact of summation between individuals who showed N400 effect and P 600 effects (Tanner & Van Hell, 2014).

[32] Note that including Region 3 did not have a large impact on the results.

task. The reliability index for the T-AGJT was .67, which is in the acceptable range. Split-half reliability, with the Spearman-Brown correction, was .52, which is acceptable in light of other studies of implicit learning (Dienes, 1992; Kaufman et al., 2010; Reber, Walkenfeld, & Hernstadt, 1991). The reliability for the visual-world task was extremely low, .02. To the best of our knowledge, no standard procedure for computing reliability existed for the visual-world task. The eye-movement data may not be consistent across trials as the looks to the pictures in the quadrant could be influenced by many other factors (e.g., attracted to interesting pictures). Given the relatively new technology of the measure, other ways of estimating reliability are needed in future research (e.g., test-retest reliability). More detailed discussion about the reliability estimation for implicit knowledge measures are presented in Appendix F.

Table 25. *Descriptive Statistics for the Language Tests*

	N	Possible Max	*M*	*SD*	Min	Max	Reliability
Eye	100	-	0.01	0.09	-0.26	0.24	.02[a33]
WM	100	-	22.10	54.47	-110.52	161.66	.91[a]
SPR	100	-	35.99	90.47	-198.47	351.27	.96[a]
T-AGJT	100	100	43.43	12.12	14.58	76.19	.67[a]
T-VJGT	100	100	30.64	16.28	2.08	82.74	.85[a]
T-SPOT	99	100	27.13	23.37	0.00	91.67	.95[a]
SRT	100	-	17.36	16.55	-35	71	.52[b]
Letter Span	100	126	89.87	14.61	52	121	.92[a]
LLAMA F	95	30	23.18	4.19	11	30	.88[a]

a. Cronbach's alpha; b. Split-half reliability, corrected using Spearman-Brown formula

Note. Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT

6.3.2 Missing Data and Data Transformation. Out of 100 participants, only one participant had missing cases for CFA and MTMM analysis, due to experimenter error in T-

[33] Since there were missing cases for the proportion of looks in the visual-world task due to eye-movement track loss, these random missing values were imputed by the Markov Chain Monte Carlo method (MCMC) for reliability estimation.

SPOT. Since this person was the only one who had a missing case in the language tests, this person was excluded list-wise from the analysis. For the SEM analyses, five participants also had missing cases for LLAMA F because they did not follow or understand the instructions. Combined with one participant who had a missing case in the T-SPOT, these random missing scores were also deleted list-wise; the total number of participants left was 94. Note that the direct ML (maximum likelihood) estimation or full information maximum likelihood (FIML) is usually recommended for treating missing cases (Brown, 2006), but the list-wise deletion procedure was chosen because there were only 6 missing cases and it allows for a wider variety of fit indices in the LISREL program.[34]

In order to compute the sum of the three indices for each target structure in the visual-world task, the indices were transformed to standardized z scores and averaged. For the RT measures (the word-monitoring and the self-paced reading tasks), the RT difference scores across the three scores were averaged after the difference for the RTs of grammatical and ungrammatical items (z scores) was computed. The z standardization controlled for the baseline RT differences among L2 learners (Faust, Balota, Spieler, & Ferraro, 1999), treating the sensitivity across the target structures equally.

Prior to the CFA and MTMM analyses, tests of univariate normality were examined for the six test scores. The total scores of the T-SPOT were positively skewed; square root transformation was applied to reduce skewness. Based on the standardized coefficients of skewness and kurtosis (*z* scores), all the variables met the assumption of univariate normality ($p > .05$). Multivariate normality of the score distribution for all the 6 variables was examined by

[34] The results obtained with the direct ML estimation did not change the overall pattern of results.

Mardia's coefficient. The coefficient (chi-square) was 1.648 (p = .439), which met the assumption of multivariate normality.

For the SEM analyses, score distribution of an additional three indicators (SRT, Letter Span, and LLAMA F) were examined. The distribution of LLAMA F scores was negatively skewed, so it was transformed to reduce the skewness via inverse transformation. The assumption of univariate normality was met for all three variables (p > .05). Assumption of multivariate normality for all 9 variables was also met; Mardia's coefficient was .014 (p = .993).

6.4 Construct Validation of Explicit and Implicit Knowledge Measures

6.4.1 Confirmatory Factor Analysis. Pearson's correlation coefficients were first computed among the six language test scores. The three competing models were submitted to CFA. The model was first tested with the whole group, followed by the short-LOR group and the long-LOR group.

***6.4.1.1 Whole Group*.** Table 26 shows the correlation matrix for the six linguistic test scores in the L2 speakers. Significantly moderate relationships were found among the timed form-focused tasks (.508 < r < .681), whereas the correlations among the three online tests were weak, and the only significant relationship among the online measures, between the word-monitoring and the self-paced reading tasks, was weak (r = .261, p = .009). The visual world task was only significantly correlated with T-SPOT.

Table 26. *Intercorrelations of the Language Tests (Whole Group, n = 99)*

	Eye	WM	SPR	T-AGJT	T-VGJT	T-SPOT
Eye	-	.093	.129	.153	.185	.212*
WM		-	.261**	.060	-.074	.057
SPR			-	.164	.073	.102
T-AGJT				-	.681**	.508**
T-VGJT					-	.553**
T-SPOT						-

Note. * p < .05, ** p < .01

All the three hypothesized CFA models were entered into the CFA analyses. Two correlated errors were imposed initially between the two measures that used similar methods. Since the timed visual GJT and the time auditory GJT used the same procedure except for the modality difference, it is reasonably assumed that the measurement errors between them would correlate to each other. Similarly, the word-monitoring task and the self-paced reading task both focused on the registration of errors during comprehension processing, the measurement errors were expected to be correlated. For parsimony, only the correlated errors that were statistically significant were retained in the final model. Only the correlated error between the word-monitoring task and the self-paced reading task was statistically significant and retained. The obtained model fit indices based on the maximum-likelihood estimation for the CFA models are summarized in Tables 27 and 28.

Both two-factor model (Model 1) and one-factor model (Model 2) produced a good fit, but the written and aural model (Model 3) did not converge due to the model misspecification. A chi-square difference test was conducted to compare Model 1 and Model 2; the difference was not significant, $\chi_{dff} = 0.897$, $df = 1$, $p = .344$. This suggests that both two-factor model and one-factor models are plausible for the obtained data.

Table 27. *CFA Model Fit Indices (Whole Group, n = 99)*

Model #	1	2	3
Description	Two-factor (Corr. Err.)	One-factor (Corr. Err.)	Written and aural
Model df	7	8	
$\chi 2$	6.043	6.94	
P	.535	.543	
NNFI	1.019	1.018	
CFI	1	1	Improper solution
SRMR	0.036	0.044	
RMSEA	0	0	
RMSEA 90% CI	0.0-0.113	0.0-0.107	

AIC	-919.179	-920.282

Note. NNFI = non-normed fit index, CFI = comparative fit index, RMSEA = root mean-square error of approximation, SRMR = standardized root mean square, AIC = Akaike Information Criterion.

Table 28. *CFA Model Fit Decisions (Whole Group, n = 99)*

Model #	1	2	3
Description	Two-factor (Corr. Errors)	One-factor (Corr. Errors)	Written and aural
NNFI & CFI (> .96)	✓	✓	
SRMR (< .09)	✓	✓	
RMSEA (< .06)	✓	✓	
Model Fit Decision	Good fit	Good fit	Poor fit

Figure 24 presents the final model with factor loadings and correlated errors. In Model 1, the two latent factors were moderately correlated (r = .47, p = .069). Factor loadings for AEK were high and significant, whereas those for IK were much lower and the path to EYE (the visual-world task) was only marginally significant. Latent construct reliability was assessed with a coefficient H, computed from the standardized factor loadings (Hancock & Mueller, 2001). The reliability coefficient for the AEK factor was satisfactory (H = .836), while for the IK factor it was very low (H = .282). Similarly, the factor loadings for IK were much lower than those for AEK in Model 2. The reliability coefficient for the IK factor was (H = .874).

Comparing the factor loading in Models 1 and 2, the magnitudes were the same for the AEK latent factor side. In contrast, all the factor loadings for IK were greater in Model 1. This partially supported that the two-factor model was more plausible than the one-factor model.

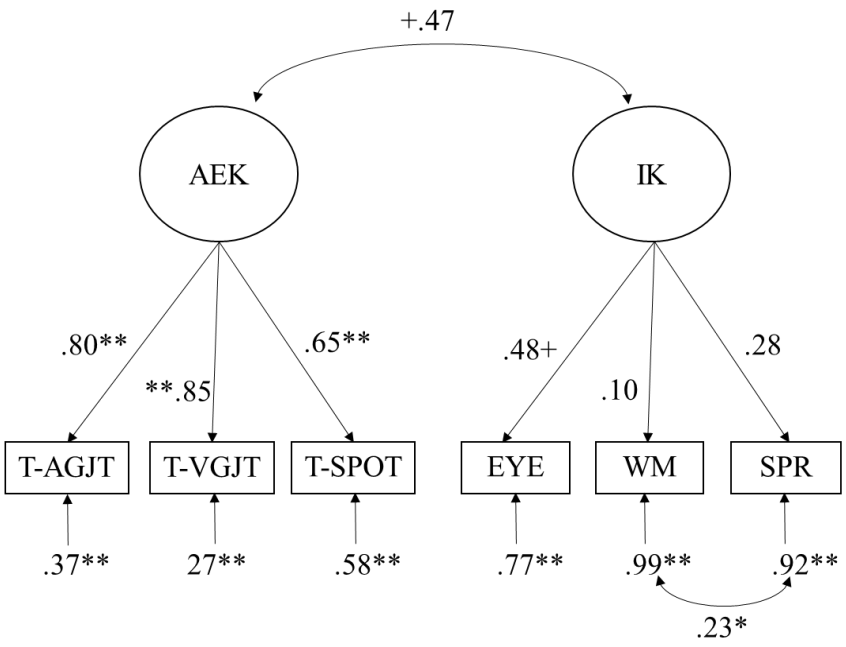


Figure 24. CFA Model 1: Two-Factor Model (Whole Group, n = 99)
Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

Note. +Standardized coefficient $p < .10$, *Standardized coefficient $p < .05$, **Standardized coefficient $p < .01$

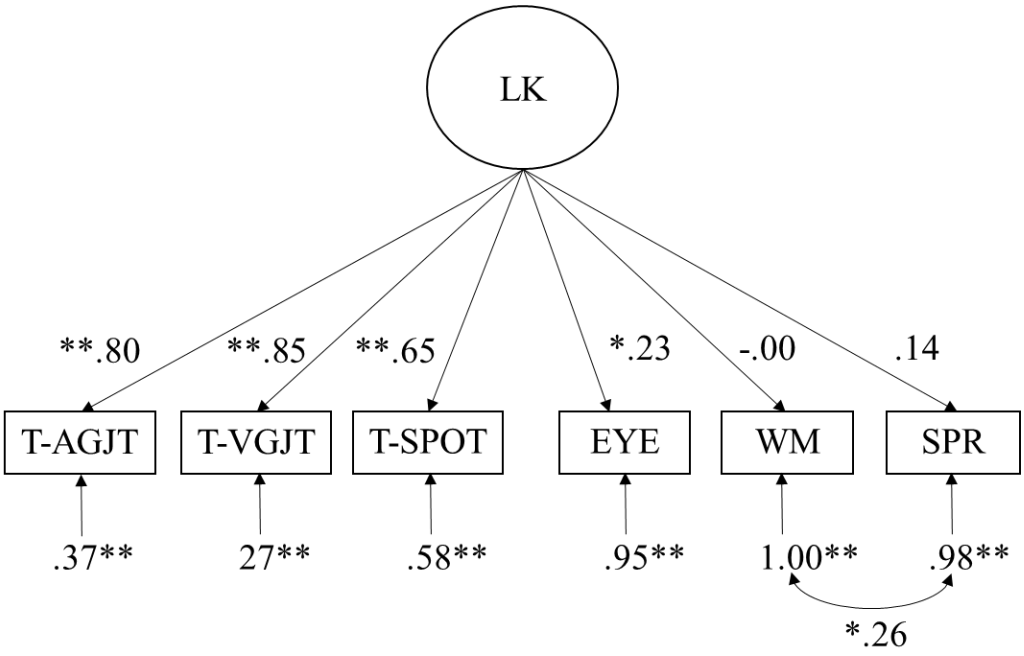


Figure 25. CFA Model 2: One-Factor Model (Whole Group, n = 99)
Note. IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

Note. +Standardized coefficient $p < .10$, *Standardized coefficient $p < .05$, **Standardized coefficient $p < .01$

6.4.1.2 Short-LOR group. In order to investigate how the amount of L2 experience, indicated by LOR, changes the validity of the test battery, CFAs were conducted separately for the two subsets. The correlation matrix among the measures for the short-LOR group is presented in Table 29. The form-focused tasks converged to a similar extent as the whole group ($.515 < r < .691$); however, no meaningful relationships were found among the three online tasks.

Table 29. *Intercorrelations of the Language Tests (Short-LOR Group, n = 47)*

	Eye	WM	SPR	T-AGJT	T-VGJT	T-SPOT
Eye	-	-.129	-.057	.146	.217	.170
WM		-	.100	.130	-.010	.165

SPR	-	.142	.137	.128
T-AGJT		-	.691**	.515**
T-VGJT			-	.539**
T-SPOT				-

The same three CFA models were evaluated for the short-LOR group (Tables 30 and 31). For the one-factor model, any of the correlated errors were not significant and the model without any correlated errors was retained as a final model. The three criteria met only for the one-factor model. The two-factor model failed to converge, regardless of the correlated errors added. The written and aural model also resulted in improper solutions. As in the whole group analysis, loadings for AEK were sufficiently high, but the loadings for the IK factor were lower than .3. The construct reliability was high (H = .841).

Table 30. *CFA Model Fit Indices (Short-LOR group, n = 47)*

Model #	1	2	3
Description	Two-factor	One-factor	Written and aural
Model df	8	9	8
$\chi 2$	Improper solution	4.894	Improper solution
p		.844	
NNFI		1.159	
CFI		1	
SRMR		0.055	
RMSEA		0	
RMSEA 90% CI		0.0-0.094	
AIC		-427.924	

Note. Both Model 3 (Written and Aural Model) and Model 3a (Written and Aural Model with correlated errors) resulted in improper solution.

Table 31. *CFA Model Fit Decisions (Short-LOR group, n = 47)*

Model #	1	2	3
Description	Two-factor	One-factor	Written and aural
NNFI & CFI (> .96)		✓	
SRMR (< .09)		✓	
RMSEA (< .06)		✓	

Model Fit Decision	Poor fit	Good fit	Poor fit

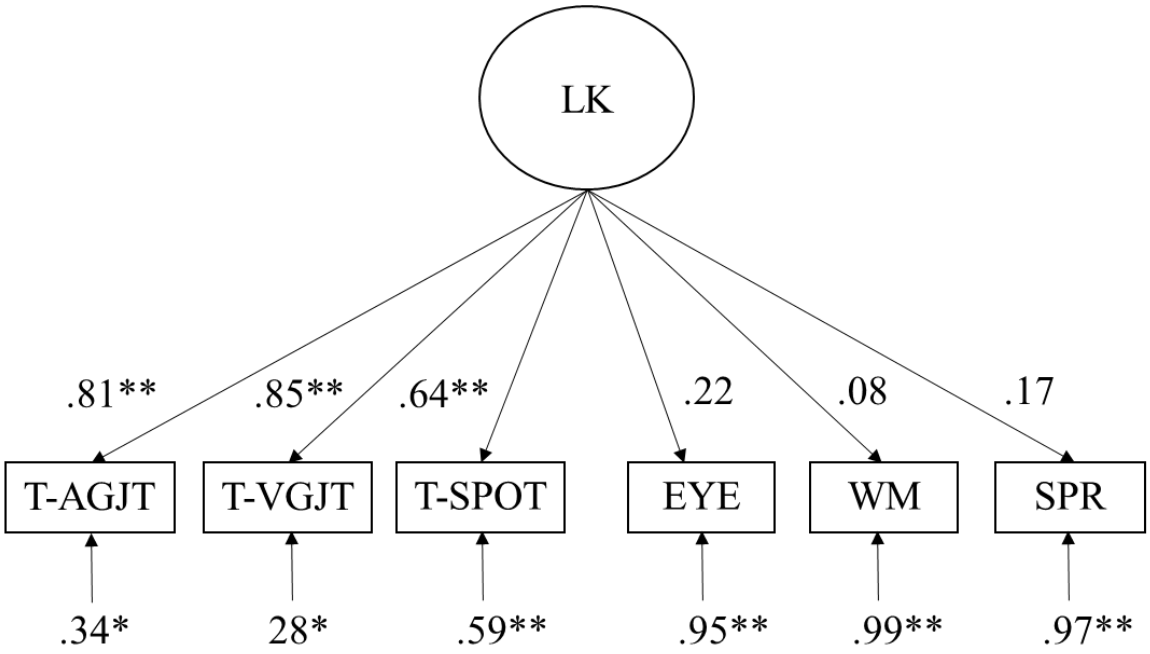


Figure 26. CFA Model 2: One-Factor Model (Short-LOR Group, n = 47)
Note. LK = Linguistic Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

Note. +Standardized coefficient $p < .10$, *Standardized coefficient $p < .05$, **Standardized coefficient $p < .01$

6.4.1.3 Long-LOR group. The correlation matrix among the language measures for the short-LOR group (Table 32). The form-focused tasks converged to a similar extent as the whole group ($.534 < r < .626$). The three online measures were more correlated with each other more in the long-LOR group than in the whole group ($.237 < r < .343$).

Table 32. *Intercorrelations of the Language Tests (Long-LOR Group, n = 52)*

	Eye	WM	SPR	T-AGJT	T-VGJT	T-SPOT
Eye	-	.237	.343*	.158	.131	.266
WM		-	.270	.010	-.157	.018

SPR	-	.173	.012	.077
T-AGJT		-	.626**	.534**
T-VGJT			-	.567**
T-SPOT				-

For the long-LOR group, the two-factor model without correlated errors and the one-factor model with the correlated error of the word-monitoring task and the self-paced reading task were retained as final models. The written and aural model resulted in an improper solution.

Inspection of model fits showed that none of the fit indices reached the criteria in the one-factor model, indicating a poor fit (Table 33). In contrast, the two-factor model produced a good fit, and the factor loadings are presented in Figure 27. As in the whole group model, factor loadings from AEK were consistently high and the reliability estimate for the AEK was also high (H = .811). For the IK factor, factor loadings were higher than in the model for the whole group, and they were all significant. The construct reliability was moderate (H = .567). Interestingly, the covariance between AEK and IK was lower in the long-LOR group (r = .22, p = .258), suggesting that the two latent factors were more distinct in the long-LOR group than the whole group.

Table 33. *CFA Model Fit Indices (Long-LOR group, n = 52)*

Model #	1	2	3
Description	Two-factor	One-factor (Corr. Errors)	Written and aural
Model df	8	8	Improper solution
$\chi 2$	7.527	13.104	
p	.481	.108	
NNFI	1.015	0.833	
CFI	1	0.911	
SRMR	0.071	0.099	
RMSEA	0	0.111	
RMSEA 90% CI	0.0-0.156	0.0-0.215	
AIC	-541.139	-535.562	

Note. The written and aural models (with and without the correlated errors) resulted in improper solution.

Table 34. *CFA Model Fit Decisions (Long-LOR group, n = 52)*

Model #	1	2	3
Description	Two-factor	One-factor (Corr. Errors)	Written and aural
NNFI & CFI (> .96)	✓		
SRMR (< .09)	✓		
RMSEA (< .06)	✓		
Model Fit Decision	Good fit	Poor fit	Poor fit

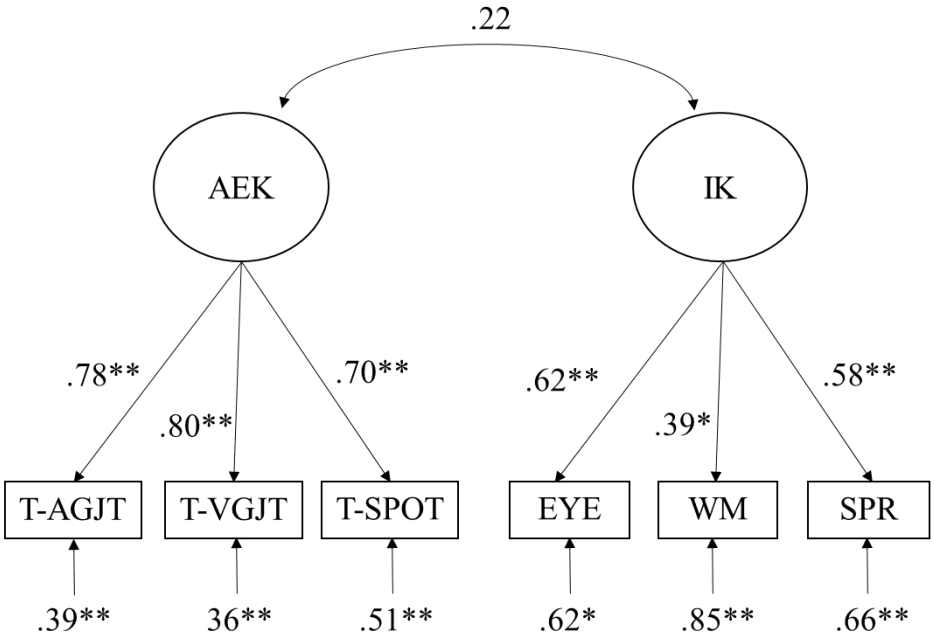


Figure 27. CFA Model 1: Two-Factor Model (Long-LOR Group, n = 52)

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

Note. +Standardized coefficient $p < .10$, *Standardized coefficient $p < .05$, **Standardized coefficient $p < .01$

6.4.2 Multi-Trait Multi-Method Analysis. The fit indices of the correlated uniqueness model indicated a good fit, χ^2 (9, N = 99) = 9.06, p = .432, with all the three types of acceptable fit indices: absolute fit (SRMR = 0.0702), incremental fit (NNFI = 0.999 and CFI = .999), and fit for parsimony (RMSEA = 0.008, 90% Confidence Interval = 0-0.114, AIC = -920.162). The model results showed that all the trait (factor) loadings were statistically significant ($p < .05$). As in the CFA models, the factor loadings were moderate to large in the automatized explicit knowledge measures (range = .55-.93), whereas the trait loadings for the implicit knowledge were small to moderate (range = .23-.44). The construct reliability was high for the AEK factor (H = .881) and low for the IK factor (H = .279). These findings lend support for the convergent validity, particularly for the automatized explicit knowledge. A small and non-significant correlation between the two traits was found (r = .32, p = .175), which provides a good piece of evidence for discriminant validity.

The presence of method effects was examined by the correlated uniqueness (errors) among the similar methods. Although the correlated uniqueness was significant between the visual GJT and the auditory GJT (r = .36, $p < .001$), its magnitude was smaller than any of the trait (factor) loadings from the two GJTs (.55 and .59). The correlated uniqueness between the word-monitoring task and the self-paced reading task was not significant (r = .23, p = .364), and its magnitude was also smaller than any of the trait loadings (.44 and .29). Method effects evaluated in the MTMM model are marginal, indicating that the set of measurements estimated traits validly with little influence from the method effects.

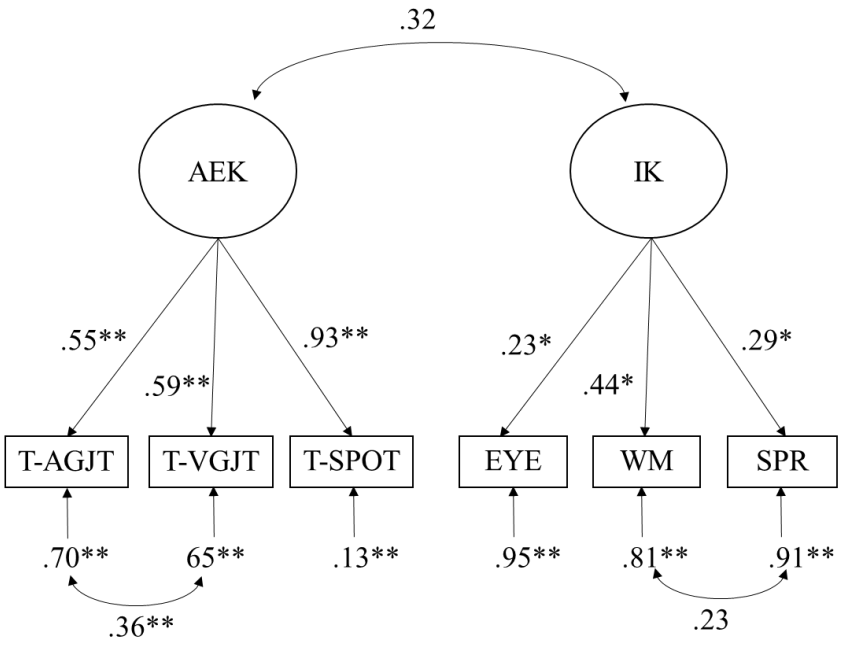


Figure 28. MTMM Model 1: Correlated Uniqueness model (Whole Group, n = 99)

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

Note. +Standardized coefficient $p < .10$, *Standardized coefficient $p < .05$, **Standardized coefficient $p < .01$

The same analysis was conducted to the short-LOR group and the long-LOR group, respectively. The model resulted in an improper solution for the short-LOR; the model for the long LOR-group is only presented (See Figure). The fit indices of the correlated uniqueness model specification indicated a good fit of the model, χ^2 (9, N = 52) = 10.622, p = .303, with most of the three types of acceptable fit indices: absolute fit (SRMR = 0.0974), incremental fit (NNFI = 0.953 and CFI = .972), and fit for parsimony (RMSEA = 0.056, 90% Confidence Interval = 0-0.173, AIC = -540.044). Although the SRMR and NNFI indices did not pass the criteria, they were close to the criteria. The overall assessment of the fit was deemed acceptable.

The model results showed that all the trait factor loadings were statistically significant ($p < .01$). The magnitude of the trait loadings was medium to large, both for the automatized explicit knowledge measures (range = .63-.86) and for the implicit knowledge (range = .40-.74). The construct reliability was high for the AEK factor ($H = .81$) and acceptable for the IK factor ($H = .621$). The analysis from the long-LOR group offered stronger convergent validity evidence for both traits than the whole-group analysis. A non-significant negligible correlation between the two traits also constitutes evidence for discriminant validity ($r = .10$, $p = .175$).

The presence of method effects was investigated through the correlated uniqueness among the similar methods. Although the correlated uniqueness was significant between the visual GJT and the auditory GJT ($r = .21$, $p < .001$), the magnitude was smaller than the trait factor loadings from the two GJTs (.63 and .66, $p < .001$). The correlated uniqueness between the word-monitoring task and the self-paced reading task was not significant ($r = -.09$, $p = .364$), and the magnitude of the trait loadings was larger than the correlated uniqueness (.44 and .74, $p < .001$). Method effects estimated in the long-LOR group were smaller than the whole group, providing support for the stability of traits.

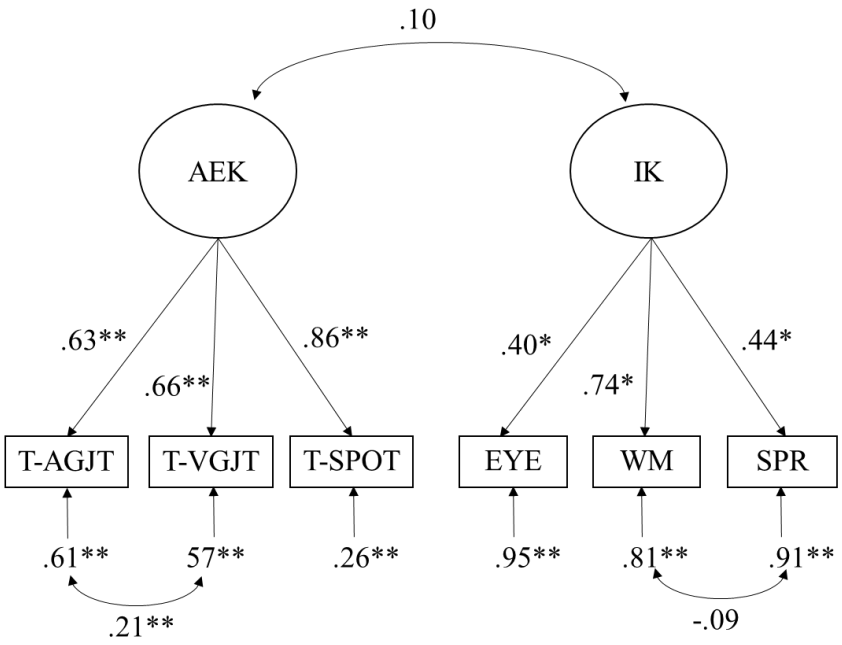


Figure 29. MTMM Model 1: Correlated Uniqueness model (Long-LOR Group, n = 52)

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

Note. +Standardized coefficient $p < .10$, *Standardized coefficient $p < .05$, **Standardized coefficient $p < .01$

6.4.3 Structural Equation Modeling Analysis. As a final step for the validation of the measures for explicit and implicit knowledge, Structural Equation Modeling (SEM) analyses were conducted. Specifically, the constructs underlying the two factors estimated by the CFA and MTMM analysis were scrutinized by investigating the nature of two types of linguistic knowledge in relation to cognitive aptitudes for explicit and implicit learning, the correlation matrix of the language measures and the aptitude measures (Table 35). The letter-span task was not used for the current SEM analysis because the primary focus is on an explicit and implicit dimension (see the next section).

The correlation coefficients among the three aptitude tests show that no relationship existed between the SRT score and the LLAMA F score. The letter-span score was weakly related to the SRT score and the LLAMA F score. When comparing the aptitude scores with the language tests, the scores on the explicit knowledge tests were weakly correlated with LLAMA F scores ($.188 < r < .235$). The SRT score was related to the eye-tracking scores more than to the other language tests ($r = .200$, $p = .053$)

Table 35. *Intercorrelations of the Language Tests and Aptitude Tests (n = 94)*

	1	2	3	4	5	6	7	8	9
1. Eye	-	.112	.120	.167	.195	.226*	.200	-.007	-.089
2. WM		-	.257*	.061	-.090	.047	.074	.050	.095
3. SPR			-	.157	.055	.084	.049	-.014	.063
4. T-AGJT				-	.684**	.505**	.089	-.015	.196
5. T-VGJT					-	.552**	.152	-.097	.188
6. T-SPOT						-	.107	.097	.235*
7. SRT							-	.167	.008
8. L-SPAN								-	.207*
9. LLAMA F									-

Note. * $p < .05$, ** $p < .01$

The SEM model was constructed by adding the two aptitude factors to the two-factor CFA model. Since the two correlated errors were statistically significant, the model with the correlated errors was retained as a final model. Table 36 shows that the model fit indices and the fit indices were good.

Table 36. *SEM Model 1 Fit Indices (Whole group, n = 99)*

Model #	1
Description	Validation Model (Corr. Errors)
Model *df*	14
$\chi 2$	8.532
p	.860
NNFI	1.096
CFI	1
SRMR	0.040

RMSEA	0
RMSEA 90% CI	0.0-0.054
AIC	-191.635

In Figure 30, the full structural equation model with correlated errors is presented with the resulting standardized coefficients. The model included two statistically significant direct effects at the .05 alpha level: the path from implicit learning aptitude to implicit knowledge (r = .21, z = 2.005, p = .045) and the path from explicit learning aptitude to automatized explicit knowledge (r = .33, z = 2.278, p = .023). The path from implicit aptitude to AEK was not significant (r = .15, z = 1.295, p = .195), nor was the path from explicit aptitude to IK (r = -.11, z = - 0.877, p = .381). These dissociation patterns further corroborate that the two factors extracted from the CFA analyses above are indeed explicit and implicit knowledge.

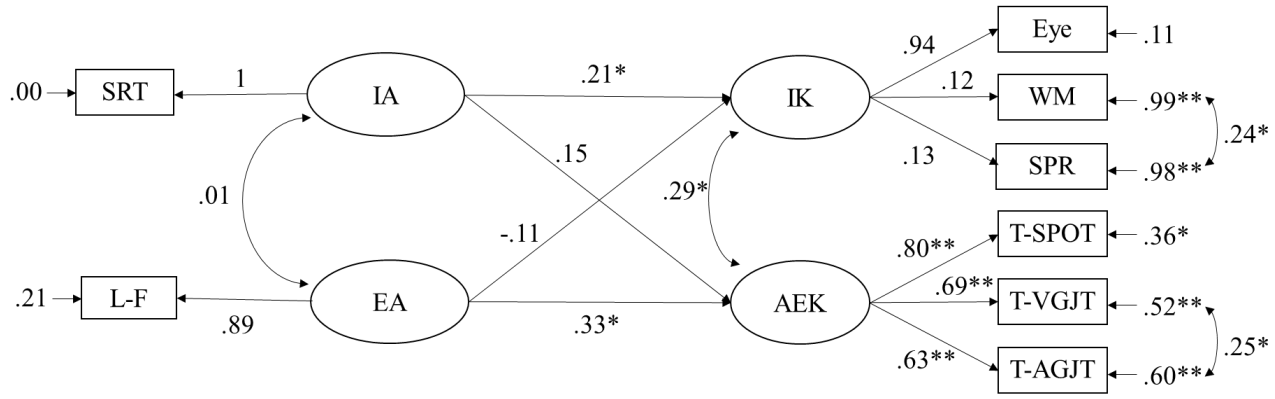


Figure 30. SEM Model 1: Validation Model (Whole Group, n = 94)

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, IA = Implicit Aptitude, EA = Explicit Aptitude, T-AGJT = Timed Auditory GJT,T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task

Note. +Standardized coefficient p < .10, *Standardized coefficient p < .05, **Standardized coefficient p < .01

6.5 The Interface Issue of Explicit and Implicit Learning and Knowledge

In order to explore the interface issue of explicit and implicit learning and knowledge, two competing models were constructed: SEM 2 (AEK to IK) and SEM 3 (no interface between AEK and IK). Fit indices for the two models are presented in Table 37. The model with the correlated errors was examined for model fit. Although the correlated error between the two GJTs was close to statistically significant at .05 (p = .066), it was retained in the final model.

As shown in Table 37, all the fit indices indicated a good fit to the data for both Models 2 and 3. A chi-square difference test was conducted to statistically compare Models 2 and 3. The models with paths between automatized explicit knowledge and implicit knowledge were significantly better than the model without the relationship between them, $\chi_{dff} = 6.087$, $df = 1$, $p = .014$. Therefore, only the results from Model 2 are reported below.

Table 37. *Summary of Fit Indices for SEM Analyses (Whole group, n = 99)*

Model #	2	3
Description	AEK to IK	No interface
Model df	18	19
$\chi 2$	13.411	19.498
p	.767	.425
NNFI	1.08	0.992
CFI	1	0.996
SRMR	0.0427	0.0666
RMSEA	0	0.017
RMSEA 90% CI	0.0 ; 0.065	0.0 ; 0.092
AIC	412.452	416.497

In Model 2 (AEK to IK), two significant effects were identified: the path from explicit learning aptitude to automatized explicit knowledge ($r = .35$, $z = 2.385$, $p = .017$) and the path from automatized explicit knowledge to implicit knowledge ($r = .33$, $z = 2.333$, $p = .020$). The

effect of explicit learning aptitude on implicit knowledge was inverse and not significant (r = -.21, z = -1.649, p = .099). The path from implicit learning aptitude to implicit knowledge was significant in the SEM Model 1, but it was not in this full SEM 2 (r = .15, z = 1.448, p = .148). In the reduced SEM model in which the relationship between implicit knowledge and automatized explicit knowledge was excluded, the path from implicit learning aptitude to implicit knowledge turned out to be significant (z = 2.002, p = 0.045).

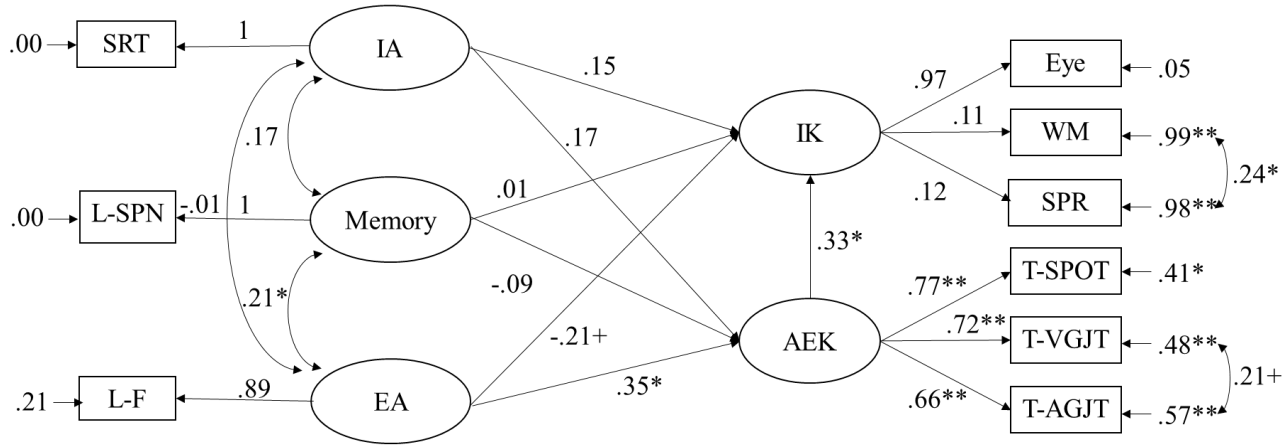


Figure 31. SEM Model 2: AEK to IK model (Whole Group, n = 94)

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, T-AGJT = Timed Auditory GJT, T-VGJT = Timed Visual GJT, T-SPOT = Timed SPOT, Eye = Visual-World task, SPR = Self-Paced Reading task, WM = Word-Monitoring task, IA = Implicit Aptitude, Memory = phonological short-term memory, EA = Explicit Aptitude, SRT = Serial-Reaction Time task, L-SPAN = Letter-span task, L-F = LLAMA F

Note. +Standardized coefficient p < .10, *Standardized coefficient p < .05, **Standardized coefficient p < .01.

Chapter 7: Discussion

7.1 Validation of Explicit and Implicit Knowledge Measures

The first research question addressed whether the three online psycholinguistic measures tap the distinct construct from the other time-pressured form-focused tests. For these measurements, six hypotheses were evaluated to examine the construct validity of automatized explicit knowledge and implicit knowledge. These hypotheses were examined first for the whole-group analysis, followed by the subset analysis to investigate whether the results would change depending on the amount of L2 exposure (research question 2).

7.1.1 Whole-Group Analysis. Table 38 provides a summary of the good-fit models identified in the current study. The reliability indices and the correlation coefficients between the two factors are presented.

Table 38. *Summary of Findings for Construct Validation of Automatized Explicit Knowledge and Implicit Knowledge Measures*

Group	Analysis	Results	*H* (AEK)	*H* (IK)	Corr. (AEK and IK)
Whole	CFA	Two-factor model	.836	.282	.47
	CFA	One-factor model			
	MTMM	Traits > Methods	.881	.279	.32
	SEM	IA -> IK, EA -> AEK	.770	.884	.30
Short-LOR	CFA	One-factor model			
Long-LOR	CFA	Two-factor model	.811	.567	.22
Long-LOR	MTMM	Traits > Methods	.810	.621	.10

Note. AEK = Automatized Explicit Knowledge, IK = Implicit Knowledge, IA = Implicit Aptitude, and EA = Explicit Aptitude. No information is provided for the one-factor models because no indices were obtained for the two factors.

The results of CFA demonstrated that the two-factor model fit the data well in the current study (Hypothesis 1). With regard to Hypothesis 2, although the factor loadings for automatized explicit knowledge were high and statistically significant (range: .65 to .85), the loadings for

implicit knowledge were much lower (range: .10 to .48) in CFA. This is also confirmed by the construct reliability index *H*; the coefficient was reliable for automatized explicit knowledge (.836 in CFA .881 in MTMM analysis), but not for implicit knowledge (.282 in CFA and .279 in MTMM analysis). The very low reliability index underscores the challenges to devise measurements for implicit knowledge. Research in psychology has repeatedly found that the reliability of the tasks used in implicit learning paradigms is much lower than that for other cognitive tasks (e.g., Reber, Walkenfeld, & Hernstadt, 1991). It seems that devising implicit tasks for linguistic knowledge is as difficult as or probably even harder than devising domain-general implicit learning tasks. The evidence indicates weak convergent validity for the measurements for implicit knowledge.

In contrast, supporting evidence was provided for the discriminant validity (Hypotheses 3a and 3b). The fit of the one-factor model (i.e., the perfectly-correlated factor) was as good as the two-factor model. However, the correlation between the two factors in the two-factor model was not strong and non-significant, both in the results of the CFA (r = .47, p = .069) and in the MTMM analysis (r = .32, p = .175). Given that "a factor correlation that exceeds .80 or .85 is often used as the criterion to define poor discriminant validity" (Brown, 2006, p. 131), the results support that the two factors were extracted successfully. In addition, the factor loadings from the three online measurements were all lower in the one-factor model than in the two-factor model, suggesting that the two-factor model can account for the correlation matrix of the current data set better than the one-factor model.

The MTMM analysis further assessed the method effects that stemmed from the artifact of having similar methods for grammaticality judgment tests and also for tests based on reaction time techniques (Hypothesis 4). The results showed that the correlated error of the two GJTs was

significant but smaller than the trait factor loadings. The correlated error of the word-monitoring task and the self-paced reading task was not significant and the coefficient was lower than the trait factor loadings. Although method effects exist to some extent, the traits were measured reliably. Note that the factor correlation was smaller when accounting for the method effects, estimating the *true* covariance of automatized explicit knowledge and implicit knowledge.

Last, the nature of the two latent constructs was examined through the cognitive aptitudes for explicit and implicit learning (Hypotheses 5-6). Results showed that the latent factor hypothesized to be implicit knowledge was significantly predicted by the SRT score (aptitude for implicit learning), but not by the explicit learning aptitude. In contrast, the latent factor hypothesized as automatized explicit knowledge was significantly predicted by the LLAMA F score (aptitude for explicit learning), but not by the SRT score. The dissociative relationship between the cognitive aptitudes and the forms of linguistic knowledge lends strong support to the validity of the explicit and implicit knowledge measures.

The significant zero-order correlation between the SRT scores and the index from the word-monitoring task has been found in previous studies (Granena, 2013b; Suzuki & DeKeyser, in press), and the current findings further corroborated this through evidence at the latent construct levels in SEM analysis. The magnitude of the path coefficient from implicit learning aptitude to implicit knowledge was lower (r = .21, p = .045) than that from explicit learning aptitude to explicit knowledge (r = 34, p = .023). This could be attributed to the instability of implicit knowledge measures, as shown in lower convergence in those measures. It is emphasized that capturing the implicit processes and implicit knowledge is not impossible but extremely difficult.

To sum up the construct validation process in the whole-group analysis, all the hypotheses were supported except for the convergent validity of implicit knowledge. The lack of convergent evidence particularly for implicit knowledge (e.g., the low factor loadings) highlights the challenges in tapping into implicit knowledge reliably. Not all late L2 learners are able to use implicit knowledge consistently, even if the tasks are designed to draw on implicit knowledge and shut off the access to (automatized) explicit knowledge. Despite the unstable nature of implicit knowledge and its measurements, some evidence like the positive correlation with the implicit learning aptitude scores support the claim that the "implicit knowledge measures" are purer measures of implicit knowledge than the measures of automatized explicit knowledge. The low convergence issue could be probed further for its sources. Although implicit knowledge could be just extremely hard to reliably assess due to its nature, it may be the case that more experienced L2 speakers, who tend to rely more on implicit knowledge, could deploy implicit knowledge more stably than less experienced speakers. This assumption was empirically tested by dividing the group based on LOR, which answered the second research question and tested Hypothesis 7.

7.1.2 Subset Analysis. Results of subset analysis demonstrated a different pattern for the two L2 groups differing in the amount of L2 experience indexed by LOR. For the short-LOR group, the two-factor model did not converge, but the one-factor model produced a good fit. In contrast, the two-factor model, but not the one-factor model, fit the data well for the long-LOR group. This suggests that L2 speakers with more experience can reliably deploy the implicit knowledge measured with the three psycholinguistic measures. L2 speakers with less experience, however, do not seem to be able to deploy implicit knowledge consistently, as indicated by the low factor loadings for implicit knowledge.

Inspecting the results from the two-factor model in the long-LOR group, the factor loadings for automatized explicit knowledge were as good as for the whole group (range: .70 to .80). More importantly, the factor loadings for implicit knowledge were higher and statistically significant; the two moderate loadings (.58 for the self-paced reading task and .62 for the visual-world task) and one weak loading (.39 for the word-monitoring task) in CFA. The *H* coefficients indicated good reliability for automatized explicit knowledge (.811 in CFA and .810 in MTMM analysis) and acceptable reliability for implicit knowledge (.567 in CFA and .621 in MTMM analysis).

Further support for discriminant validity was also obtained for the long-LOR group: The correlation between the two factors was non-significant and weaker than for the whole group in the CFA (r = .22, p = .258) and in MTMM analysis (r = .10, p = .175). The low correlation between the factors can be interpreted as both automatized explicit knowledge and implicit knowledge having been assessed distinctively through the six measures. The greater experience in the immersion context seemed to have made L2 speakers rely more on implicit knowledge, and it resulted in the more consistent and stable use of implicit knowledge.

The MTMM analysis for the long-LOR group further corroborated that the correlated error of the two GJTs was significant but smaller than the trait factor loadings, and that that of the word-monitoring task and the self-paced reading task was a non-significant small negative value. The traits seemed to be assessed more reliably with negligible method effects.

In sum, the overall findings from the long-LOR group supported all the hypotheses more strongly, including the convergent validity of implicit knowledge. Even though implicit knowledge is much harder to assess, compared to (automatized) explicit knowledge, it is possible to tap into implicit knowledge more stably, particularly when more experienced L2 speakers

performed the test battery. This corroborated the previous findings in Suzuki and DeKeyser (in press) and is consistent with Paradis' (2009) claim that explicit and implicit knowledge coexist in the L2 system, and the reliance of implicit knowledge increases over time through more L2 experience.

The current study aimed at extending the findings in the validation study for elicited imitation in Suzuki and DeKeyser (in press), who showed that it is possible to tap into automatized explicit knowledge and implicit knowledge separately when fine-grained psycholinguistic tasks are employed to examine real-time sentence processing for comprehension. The current study, as well as Suzuki and DeKeyser (in press), cast doubt on the construct validity of the previous test battery of explicit and implicit knowledge developed by R. Ellis (2005) and further utilized by others (Bowles, 2011; Ercetin & Alptekin, 2013; Sarandi, 2015; Zhang, 2014). Although previous research is cautious in stating that timed GJTs are an impure measure for implicit knowledge (e.g., Loewen, 2007), it is emphasized that time-pressure does not guarantee the inaccessibility of automatized explicit knowledge (DeKeyser, 2003; Suzuki & DeKeyser, in press). As the current findings suggest, it may be more conceivable to consider the timed form-focused tests as measures of automatized explicit knowledge, at least for the group of instructed L2 speakers and similar L2 speakers with both instructed and immersion backgrounds. Furthermore, regardless of the three analyzed groups, the factor loadings for automatized explicit knowledge were high (range: .63 to .93), and the reliability index was also high in CFA and MTMM (range: .810 to .881). This suggests that late L2 learners with some formal instruction, as was the case for the present study, tend to rely on explicit knowledge very consistently (DeKeyser, 2007b; Paradis, 2009).

Results from the CFA in the whole group and subset analyses demonstrated that the timed visual GJT had the strongest factor loading for automatized explicit knowledge. The timed auditory GJT yielded comparable loadings. The loading for the timed SPOT, however, was consistently lower than the timed GJTs. Both GJTs included ungrammatical sentences which participants were required to consciously detect, whereas SPOT asked them to fill in the blank with the target structures. Detecting anomaly of the sentences consciously in the timed GJTs may draw more on explicit knowledge. Note that results from the MTMM analysis suggest that the correlated errors between the GJTs were statistically significant. Despite the myriads of research utilizing GJTs in the SLA fields, little research has probed to what extent the GJTs introduce measurement errors from the task characteristics. In light of the higher loadings of the two GJTs, the current findings suggest method artifacts are not detrimental to tapping into automatized explicit knowledge.

For the implicit knowledge factor, the factor loadings from the visual-world task were the largest in all the analyses. It suggests that the visual-world task is the best measurement for implicit knowledge in the current test battery. The findings can be explained by the design of the visual-world task. First, the visual world did not use any ungrammatical sentences, which takes the attention completely away from form. Second, since the linguistic processing can be time-locked within a few milliseconds, there should be minimal opportunity for listeners to apply their explicit knowledge quickly, especially given the complete lack of focus on form. Third, the visual-world task directly captured the linguistic processing via eye-movements without any mediation such as through button presses. Put differently, the visual-world task is more advantageous because there is no need to make an overt conscious decision while listening to the critical sentences. Indeed, the shared reaction time measurement technique—button responses—

in the self-paced reading task and the word-monitoring task appears to produce some measurement inaccuracies. The MTMM analysis in which the method effects of the reaction time technique were accounted for showed that the word-monitoring task loaded most strongly on the implicit knowledge factor. In other words, this suggests that the method artifacts attenuated the magnitudes of factor loadings from the reaction time measurements. All in all, it is tentatively concluded that the visual-world task is the best measure for implicit knowledge.

Having said that, the reliability index computed for the visual-world task in the current study was extremely low (.02), which should be considered to be one of the largest limitations in the study. It is counter-intuitive that the least reliable measure produced the highest loading for implicit knowledge. Since fixation data are influenced by multiple sources other than linguistic processing (e.g., picture attractiveness) across trials, it may be difficult to estimate the reliability based on internal consistency measures among the test items. It may be more suitable to estimate a test-and-retest reliability for these eye-movement data (Klein & Fischer, 2005). Furthermore, the current time-locking of the 200 ms from the post-hoc onset of the linguistic trigger needs further justification because it is possible that the reliability may improve if we include a longer region for computing sensitivity indices. Nevertheless, the current study aimed at capturing the very early sensitivity to the linguistic structures, and different sources of knowledge appeared to be recruited even within 1000 ms time windows (see Appendix E). As the application of the eye-tracking technique for assessing individuals' linguistic knowledge rather than at the group level has just begun, more research is needed to examine the reliability and validity of the eye-tracking method.

7.2 Interface of Explicit and Implicit Knowledge and Learning

Although there are some differences in the theorizing on explicit and implicit knowledge/learning and their interaction, the review of literature on the interface issue suggested that most SLA researchers agree on the two common claims:

- Explicit and implicit knowledge and learning are distinct entities that exist independently. Explicit knowledge does not turn into implicit knowledge.
- Explicit knowledge and learning play facilitative roles in the acquisition of implicit knowledge.

Based on these claims, the current study empirically investigated the extent to which automatized explicit knowledge impacts the acquisition of implicit knowledge (research question 3). Furthermore, the role of explicit and implicit learning aptitudes were investigated in order to reveal the contribution of explicit and implicit learning processes in late L2 learners (research questions 4-5). Results showed that the SEM model with a path from automatized explicit knowledge to implicit knowledge fit significantly better than the model presupposing no relationship between automatized explicit and implicit knowledge. This suggests that there is an influence of automatized explicit knowledge on implicit knowledge. For the relationship between aptitudes and linguistic knowledge, SEM analyses revealed that the only significant path was from explicit learning aptitude to automatized explicit knowledge (r = .35, p = .017). Explicit learning aptitude was negatively related to implicit knowledge (r = -21, p = .098), and implicit learning aptitude did not seem to have a systematic influence on the acquisition of either explicit or implicit knowledge. Note, however, that implicit learning aptitude does seem to influence acquisition of implicit knowledge in the reduced SEM model, in which there is no path from automatized explicit knowledge to implicit knowledge.

In summarizing the findings pertaining to research questions 3 to 5 together, to provide the broader picture, SEM Model 2 results revealed the following systematic paths: explicit learning aptitude predicted the acquisition of automatized explicit knowledge, and then automatized explicit knowledge had an impact on the development of implicit knowledge. It has been argued that explicit learning mechanisms compensate for the diminishing ability for implicit learning in adult L2 learners (DeKeyser, 2000; DeKeyser, et al., 2010), and that explicit learning aptitudes may facilitate the acquisition of implicit knowledge. Based on Bley-Vroman's Fundamental Difference Hypothesis, DeKeyser verified the hypothesis that "the only way that an adult learner can achieve grammatical competence similar to that of a native is by using analytical, problem-solving abilities, because the implicit learning mechanisms of the child are no longer available or accessible" (DeKeyser, 2000, p. 514-515). In light of the current findings, his idea can be further developed and elaborated: explicit learning processes are first primarily at work to acquire explicit knowledge and achieve more automatic use of it, which helps implicit learning mechanisms focus on some grammatical structures in the input and ultimately leads to implicit knowledge. Given the limited access to implicit learning mechanisms for adults, analytical, problem-solving abilities may be essential for most late L2 learners.

The current study recruited late L2 speakers who received classroom instruction and had at least two years of immersion experience in Japan. Overall, their general learning processes can be described as follows: explicit learning processes were first deployed in order to acquire explicit knowledge, which resulted in more automatized explicit knowledge through extensive practice. Explicit knowledge allowed L2 learners to process the language more input efficiently, e.g., attending to relevant grammatical features in input so that the implicit learning system took them in. Explicit knowledge also allowed them to frequently use the relevant grammatical

structures accurately, which also accumulated language input to implicit learning systems (N. C. Ellis, 2005, in press; Paradis, 2009).

The current study provides the first empirical evidence for the impact of automatized explicit knowledge on implicit knowledge at the latent construct level. This is consistent with the current view taken by most SLA researchers that explicit knowledge facilitates the development of implicit knowledge, refuting Krashen's idea that explicit knowledge never impacts implicit knowledge. Although no researchers have publicly claimed that explicit knowledge that was automatized should influence implicit knowledge, at least two prominent researchers, Nick Ellis and Michel Paradis, seem to agree with the idea that automatized explicit knowledge influences the acquisition of implicit knowledge (Nick Ellis, personal communication, October 24, 2014; Michel Paradis, personal communication, October 21, 2014). As the current study did not measure less- or non-automatized explicit knowledge, it is left open whether less automatized explicit knowledge has a similar impact on the development of implicit knowledge. It is speculated, however, that explicit knowledge that is at least partially automatized may be necessary for both comprehension and production because communicative interactions usually take place in real time, and explicit knowledge that can be deployed quickly should be more beneficial in tuning in to the relevant input.

Although the explicit learning processes are important for late L2 learners, explicit learning processes do not appear to directly influence the acquisition of implicit knowledge. What is needed for acquiring implicit knowledge is the product of explicit learning (i.e., automatized explicit knowledge). This is indicated by the weak negative correlation between the explicit learning aptitude and implicit knowledge (r = -21, p = .099). Since the coefficient is not statistically significant and the magnitude is small, the interpretation of this negative relationship

should be taken with caution. It can be speculated, however, that explicit learning processes could hinder implicit learning processes under some circumstances. For instance, Robinson (2005) employed an incidental (implicit) artificial grammar learning task based on Reber et al. (1991) and found a negative correlation between IQ and learning outcome. In the literature of the probabilistic SRT task performance, it appears that development of the ability for explicit learning, which usually takes place at around 11-13 years old, is related to the gradual decline in the ability to learn implicitly (Nemeth, Janacsek, & Fiser, 2013). This idea is still in the realm of speculation, but further research should examine the role of explicit learning aptitude on the development of implicit knowledge.

As was shown in the reduced SEM model (without the relationship between automatized explicit knowledge and implicit knowledge), implicit learning aptitude was a significant predictor if no path was presupposed from automatized explicit knowledge to implicit knowledge. Late L2 learners have not completely lost the capacity for learning implicitly to acquire L2 grammar. As a number of laboratory-based research studies have shown, certain L2 grammatical features can be acquired without awareness (Leung & Williams, 2011, 2012, 2014; Rebuschat, et al., 2013; Rebuschat & Williams, 2011; Williams, 2005). From the results of the current study, implicit learning routes seem to be limited, however, and a stronger learning path was explicit learning, which could proceduralize and automatize linguistic knowledge, indirectly impacting the acquisition of implicit knowledge.

The current study also investigated the role of phonological short-term memory because this basic memory ability was assumed to underlie the ability for both explicit and implicit learning and was found to predict the acquisition of L2 grammar (e.g., Martin & N. C. Ellis, 2012). The current study, however, revealed that the effects of phonological memory were

insubstantial. Since the current study examined the role of phonological short-term memory in combination with explicit and implicit learning aptitudes, this may mean that more high-order cognitive aptitudes, language inductive analytic and probabilistic sequence learning ability, play more important roles in grammar learning. Having said that, previous study found that phonological short-term memory, in combination with implicit sequence learning and associate memory abilities, predicted high-level attainment in L2 reading and listening proficiency (Linck, Hughes, et al., 2013). Since Linck et al. (2013) did not focus on the acquisition of specific grammatical structures or did not include individual difference measures of language analytic ability, it is yet to be determined whether the phonological short-term memory plays a role in L2 grammar learning. It may be the case that other components of working memory such as executive function (Engle, 2002) are related to the development of explicit morphosyntactic knowledge (Brooks, Kempe, & Sionov, 2006; Trofimovich, Ammar, & Gatbonton, 2007). To the best of our knowledge, the current study is the first attempt to examine the role of phonological short-term memory juxtaposed with explicit and implicit learning aptitudes in the learning of specific structures. Future research should examine to what extent individual differences in those three cognitive aptitudes together explain the variations in L2 attainment, both in naturalistic settings and in more controlled laboratory-based experiments.

Last but not least, the current findings bear broad implications for effective L2 instruction and learning. The fact that automatized explicit knowledge ultimately led to the acquisition of implicit knowledge underscores the value of explicit learning—deliberate practice for achieving automatized explicit knowledge (DeKeyser, 2007a). As a first step, learning solid declarative knowledge is essential for further systematic practice leading up to proceduralization and partial automatization. Through more extensive practice, full automatization and implicit knowledge

can eventually be attainable for some of the structures. Since full automatization of explicit knowledge and attainment of implicit knowledge require considerable time and effort, realistic goals for L2 classroom instruction and learning are achieving proceduralization and partial automatization, which build on initial declarative learning. The learning processes delineated above, however, may be most applicable to learners with a high level of education and ample experience with formal instruction; individual differences among late L2 learners should be also considered. Indeed, a significant contribution of explicit learning aptitudes to the acquisition of automatized explicit knowledge highlights the importance of language-analytic ability. In order to further clarify the learning processes for subsets of late L2 learners, future research needs to examine the interaction between aptitudes and instruction/learning (Cronbach & Snow, 1977).

Chapter 8: Conclusions and Future Directions

The present dissertation set out to achieve two related goals aimed at a better understanding of explicit and implicit learning systems in SLA. The first goal was to validate the more fine-grained implicit knowledge measures; this was motivated by the hypothesis from Suzuki and DeKeyser (in press) that implicit knowledge can be assessed through real-time sentence comprehension processing and resulting online registration or detection of grammatical errors. The current study generated results that are both promising and challenging for the validation of implicit knowledge. Although automatized explicit knowledge was assessed relatively easily by the conventional time-pressured form-focused tasks, it seems to be much harder to tap into implicit knowledge through behavioral measures. An array of validity evidence was provided to support that the online psycholinguistic measures successfully assessed implicit knowledge to some extent, but cautious use of the measures is recommended. In particular, as indicated by Suzuki and DeKeyser (in press), the amount of L2 experience indexed by LOR seemed to play a critical role in the stable use of implicit knowledge.

The second but ultimate goal was to empirically explore the interface issues of explicit and implicit knowledge and learning. The body of literature on the interface issue suggests a facilitative role of explicit knowledge, but little research has investigated it empirically with valid implicit knowledge measures. The present study addressed this gap and confirmed the major claims regarding the facilitative role of explicit knowledge. The present findings have further enriched the understanding of explicit and implicit learning processes in adult SLA: automatized explicit knowledge, developed through deliberate practice using explicit learning mechanisms, influences the acquisition of implicit knowledge. These findings should be

interpreted cautiously, however, because the measurements for implicit knowledge were less reliable than those for automatized explicit knowledge.

That being said, the current study opens several avenues for future research. First and foremost, more rigorous validation studies are greatly needed for developing implicit knowledge measures. Psycholinguistic measures like eye tracking to index individuals' linguistic knowledge are relatively new and unexplored for assessing individuals' linguistic knowledge in the SLA field. The index from the visual-world task had extremely low reliability. Furthermore, weaker convergent validity and large measurement errors were observed in the implicit knowledge measures. More efforts should be made to better calibrate and otherwise improve the measurements for implicit knowledge.

There were many decision points when computing indices from online measures that needed additional time-locking on the measures. For instance, the proportion of fixations during the 200 ms from the onset of the post-hoc critical region was chosen to index implicit knowledge. It may be the case that eye-movement data indicates different types of knowledge depending on different time points, as may happen in the self-paced reading task.

Another limitation of the current test battery was that it only consisted of receptive tasks that require limited productive knowledge. The previous study found that even with time pressure, it is very hard to limit access to automatized explicit knowledge in the EI task (Suzuki & DeKeyser, in press). An innovative method that can reliably elicit implicit knowledge in production should be devised, as well as other new online-processing measurements for implicit knowledge.

Furthermore, the range of target linguistic structures to be tested in implicit knowledge measures should be expanded. Four linguistic structures were tested in the current study, and

knowledge of them was successfully assessed via the visual-world task, which requires the most effort to design. It may not be possible to assess all the existing grammatical structures using the visual-world task, but a more variety of target structures should be explored in future studies. This is particularly important for investigating the interface issue, because the role of explicit and implicit learning may vary depending on the types of linguistic structures (e.g., Granena, 2013).

Last, the SEM findings pertaining to the interface issue mark the first empirical attempt to reveal that automatized explicit knowledge ultimately leads to implicit knowledge. The current findings, however, should be interpreted with caution. "Ultimately, SEM alone cannot establish causality. It can, however, provide some evidence necessary to support a causal inference" (Hair et al., 1998, p. 721). One of the requirements for establishing a causal relationship is time sequence, that is, the cause must be established prior to the effect. Since the current study only employed a cross-section design, it calls for a longitudinal research design or an intervention study. The most logical next step is to conduct longitudinal research, in which explicit/implicit knowledge and cognitive aptitudes are measured at earlier stages of L2 learning (e.g., first exposure in immersion settings) to predict the development of linguistic knowledge. Based on the current findings, it is hypothesized that explicit learning aptitude, measured at Time 1, predicts the acquisition of automatized explicit knowledge at Time 2 or later and automatized explicit knowledge further predicts the acquisition of implicit knowledge.

If one is to take an intervention approach to examine the causal relationship between explicit and implicit knowledge, it is probably best to extend the previous line of experimental or quasi-experimental studies in the 1990s, in which the effectiveness of explicit instruction for L2 acquisition was examined (e.g., DeKeyser, 1995; De Graaf, 1997; Robinson, 1997), as comprehensively summarized in the subsequent meta-analyses (Norris & Ortega, 2000; Spada &

Tomita, 2010). The body of these previous studies demonstrated that explicit L2 instruction facilitates the development of both explicit and implicit knowledge; however, it is far from conclusive on the role of explicit and implicit learning, mainly due to the difficulty in creating unambiguous opportunities for implicit learning and pure outcome measurements of implicit knowledge with little involvement of explicit learning and knowledge. Employing the measurements developed in the current study will open up a new venue for examining the interface issue from a psycholinguistic perspective.

That being said, the current study has underscored the substantial difficulty in eliciting implicit knowledge reliably, even from L2 speakers who possess advanced proficiency and were immersed in the target-language-speaking country at least for two years. It cannot be assumed that implicit knowledge can be developed and assessed during laboratory-based experiments and classroom studies of relatively short duration. Again, *longitudinal* extensive laboratory-based or classroom-based research can possibly reveal the development of explicit and implicit knowledge, if one really wants to tackle the issue of explicit and implicit knowledge and learning. Only if sufficient conditions were met for the development and elicitation of implicit knowledge would the current set of test batteries be optimally useful. Another option, which is more realistic and practical, is to focus on the developmental trajectory of explicit knowledge (i.e., proceduralization/automatization) in laboratory and classroom settings (DeKeyser, 1997; Hulstijn, Van Gelderen, & Schoonen, 2009; Lim & Godfroid, in press; Rodgers, 2011).

The current dissertation underscores the importance of the validation of fine-grained measures for implicit knowledge as well as the challenges of the validation. More empirical research for test validation should be conducted along this emerging line of psycholinguistic

investigations; valid and reliable measurements for implicit knowledge are the crux of empirical research on the interface between explicit and implicit knowledge.

Appendix A. Transitive/Intransitive Verb Pairs

Intransitive	English	Frequency	Transitive	English	Frequency
こぼれる	It spills.	6.77	こぼす	to spill	8.36
つぶれる	It is crushed.	12.31	つぶす	to crush	18.25
割れる	It is divided.	12.13	割る	to break	20.18
壊れる	It breaks.	25.42	壊す	to break	18.59
折れる	It breaks.	13.93	折る	to break	13.84
曲がる	It bends.	23.40	曲げる	to bend	12.77
汚れる	It becomes dirty.	22.26	汚す	to soil	9.13
沸く	It boils.	5.24	沸かす	to boil	4.38
混ざる	It is mixed.	4.75	混ぜる	to mix	27.41
溶ける	It melts.	16.63	溶かす	to melt	7.99
焼ける	It is burned.	15.43	焼く	to grill	55.21
燃える	It burns.	24.19	燃やす	to burn	11.21
砕ける	It smashes	5.42	砕く	to smash	6.06
破れる	It is torn.	9.98	破る	to tear	21.96
育つ	It grows up.	47.77	育てる	to raise	56.71
閉まる	It closes.	4.32	閉める	to close	16.93

Appendix B. Classifiers and Nouns

Jap. Classifier	Noun in Jap.	Noun in English	Chin. Classifier	Noun in Chin.
冊	絵本	picture-book	本/册	绘本
冊	雑誌	magazine	本	杂志
冊	辞書	dictionary	本	字典
冊	ノート	notebook	本	笔记本
匹	魚	fish	条	鱼
匹	猿	monkey	只	猴子
匹	犬	dog	只	狗
匹	ネズミ	mouse	只	老鼠
台	ピアノ	piano	架	钢琴
台	ベッド	bed	张	床
台	カメラ	camera	架/个	相机
台	携帯電話	mobile phone	个	手机
枚	ハンカチ	handkerchief	块	手帕
枚	シャツ	shirt	件	恤衫
枚	葉っぱ	leaves	片	叶
枚	鏡	mirror	面	镜子
着	服	cloth	套	穿着
着	ドレス	dress	条/件	裙子
着	コート	coat	件	外套
着	スーツ	suit	套	诉讼
羽	カラス	crow	只	乌鸦
羽	ニワトリ	chicken	只	鸡
羽	鳥	bird	只	鸟
羽	はと	pigeon	只	鸽子
足	スリッパ	slippers	双	拖鞋
足	ブーツ	boots	双	靴子
足	靴下	socks		
足	サンダル	sandals	双	凉鞋
軒	アパート	apartment	栋	公寓
軒	コンビニ	convenience store	个/间	便利店
軒	スーパー	supermarket	摊/店	超市
軒	居酒屋	Japanese tavern	个/间	酒馆

Appendix C. Counter-balancing of the Sentences for Transitive

	V-W	W-M	SPR	AGJT	VGJT
List 1					
Tran A (1-4)	G		U		G
Tran B (5-8)	G		G		U
Tran C (9-12)	-	G		U	
Tran D (13-16)	-	U		G	
Intran A (1-4)	-	G		U	
Intran B (5-8)	-	U		G	
Intran C (9-12)	G		U		G
Intran D (13-16)	G		G		U
List 2					
Tran A (1-4)	G		G		U
Tran B (5-8)	G		U		G
Tran C (9-12)	-	U		G	
Tran D (13-16)	-	G		U	
Intran A (1-4)	-	U		G	
Intran B (5-8)	-	G		U	
Intran C (9-12)	G		G		U
Intran D (13-16)	G		U		G
List 3					
Tran A (1-4)	-	G		U	
Tran B (5-8)	-	U		G	
Tran C (9-12)	G		U		G
Tran D (13-16)	G		G		U
Intran A (1-4)	G		U		G
Intran B (5-8)	G		G		U
Intran C (9-12)	-	G		U	
Intran D (13-16)	-	U		G	
List 4					
Tran A (1-4)	-	U		G	
Tran B (5-8)	-	G		U	
Tran C (9-12)	G		G		U
Tran D (13-16)	G		U		G
Intran A (1-4)	G		G		U
Intran B (5-8)	G		U		G
Intran C (9-12)	-	U		G	
Intran D (13-16)	-	G		U	

Note. VW = Visual-World Paradigm, SPR = Self-Paced Reading task, WM = Word-Monitoring task, AGJT = timed Auditory Grammaticality Judgment Test, VGJT = timed Visual Grammaticality Judgment Test

Appendix D. Counter-balancing of the Sentences for Classifiers

List 1	VW	WM	SPR	AGJT	VGJT
Classifier A (1-8)	G		G		U
Classifier B (9-16)	G		U		G
Classifier C (17-24)	-	G		U	
Classifier D (25-32)	-	U		G	
List 2					
Classifier A (1-8)	G		U		G
Classifier B (9-16)	G		G		U
Classifier C (17-24)	-	U		G	
Classifier D (25-32)	-	G		U	
List 3					
Classifier A (1-8)	-	G		U	
Classifier B (9-16)	-	U		G	
Classifier C (17-24)	G		U		G
Classifier D (25-32)	G		G		U
List 4					
Classifier A (1-8)	-	U		G	
Classifier B (9-16)	-	G		U	
Classifier C (17-24)	G		G		U
Classifier D (25-32)	G		U		G

Note 1. VW = Visual-World Paradigm, SPR = Self-Paced Reading task, WM = Word-Monitoring task, AGJT = timed Auditory Grammaticality Judgment Test, VGJT = timed Visual Grammaticality Judgment Test

Note 2. Classifier (1-16) contains all the eight types of classifiers (two nouns for each). Classifier (17-32) also contains all the eight types of classifiers (two nouns for each)

Appendix E.

Relationship between Eye-tracking measures and Other Language Tests at different time points

In order to investigate whether the index from the visual-world task may draw on different types of linguistic knowledge, the indices were computed at four different 200 ms time windows (i.e., 0-200 ms, 200-400 ms, 400-600 ms, and 600-800 ms). As shown in Table 39, the correlation coefficient of the index during the first 200 ms (Eye1) was moderately correlated with the index during the next 200 ms (Eye2), weakly correlated with the index during 400-600 ms (Eye3), and not correlated with the index during 600-800 ms (Eye4). As demonstrated in Table 24, this also confirmed that the index at earlier time window was not consistent with the index at the later time points.

Table 39. *Intercorrelations among Eye-Tracking Measures at Different Time Windows*

	Eye1	Eye2	Eye3	Eye4
Eye1	-	.598**	.330**	.162
Eye2		-	.742**	.432**
Eye3			-	.748**
Eye4				-

Note. The indices were compute during 0 to 200 ms (Eye1), 200 to 400 ms (Eye2), 400 to 600 ms (Eye3), and 600 to 800 ms (Eye4).

Furthermore, the correlations of the four eye-tracking indices with the other five language tests were computed (Table 40). Overall results showed that the indices at the later time windows were more correlated with the time-pressured form-focused tests. The correlations of the word-monitoring task with those eye-tracking indices became smaller as the time windows went later, but there did not seem to be systematic changes in the self-paced reading tasks in relation to the eye-tracking indices. The patterns may suggest that the later processing could more likely

involve conscious processing, but it was far from clear how different time points affect explicit and implicit language processing.

Table 40. *Intercorrelations of Eye-Tracking Measures and Other Language Tests*

	WM	SPR	T-AGJT	T-VGJT	T-SPOT
Eye1	.093	.129	.153	.185	.240*
Eye2	.093	.184	.210*	.340**	.248*
Eye3	.033	.159	.234*	.388**	.312**
Eye4	-.010	.153	.260**	.303**	.312**

Appendix F. Reliability Estimations for the Implicit Knowledge Measures

Since reliability issues are relatively unexplored for reaction-time and eye-tracking measures, we have attempted to estimate internal consistencies of those measures in more depth here. We first computed Cronbach's alpha for the word-monitoring and self-paced reading tasks, separately for grammatical and ungrammatical items and for each list (Table 41).

Table 41. *Cronbach's Alpha for Word-Monitoring and the Self-Paced Reading Tasks*

	Word-Monit.			Self-paced		
	All	Gramm.	Ungramm.	All	Gramm.	Ungramm.
List 1	.93	.77	.90	.97	.95	.93
List 2	.92	.85	.86	.95	.92	.90
List 3	.93	.90	.85	.96	.92	.94
List 4	.76	.59	.59	.97	.94	.93

Since the difference scores were used as dependent variables, we computed reliability indices separately for grammatical and ungrammatical items. The reliability coefficients for those difference scores are influenced by the two scores that are computed (Traub, 1994), and it was reasonable to estimate the reliability for each of them. One could estimate the reliability of the difference scores directly, but given that the design of the psycholinguistic tasks compared a different set of grammatical and ungrammatical sentences within subject (due to the counter-balancing of lists), we did not compute the reliability for the difference scores. Results showed a satisfactory internal consistency across all the measures, except for the lower reliability index for list 4. Overall, we can be confident that the internal consistency for the two reaction-time measures is high, which corroborates the high internal consistency across the whole test items reported in Table 25.

For the visual-world task, since we did not have grammatical and ungrammatical items, we estimated the reliability (Cronbach's alpha) separately for the target trials and non-target

trials and for each list (Table 42). Target trials are the ones where the target picture is referred to in the critical sentence. As found in the overall reliability estimation in Table 25, no satisfactory consistency was found in any of the conditions. This further calls for a different method of estimating reliability for the visual-world task such as test-retest reliability.

Table 42. *Cronbach's Alpha for Eye-Tracking Measures*

	All	Target	Non-Target
List 1	.09	.23	-.54
List 2	-.04	-.18	.01

References

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition, 30*(4), 481-509.

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning, 59*(2), 249-306.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247-264.

Altmann, G., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*(4), 502-518.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369-406.

Anderson, J. R. (1996). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (2005). *Cognitive psychology and its implications* (6th ed.). New York: Worth Publishers.

Andreu, L., Sanz-Torrent, M., & Trueswell, J. C. (2013). Anticipatory sentence processing in children with specific language impairment: Evidence from eye movements during listening. *Applied Psycholinguistics, 34*(1), 5-44. doi: doi:10.1017/S0142716411000592

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition, 76*(1), B13-B26.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1982). The Construct Validation of Some Components of Communicative Proficiency. *Tesol Quarterly, 16*(4), 449-465.

Batterink, L., & Neville, H. J. (2013). The human brain processes syntax in the absence of conscious awareness. *The Journal of Neuroscience, 33*(19), 8528-8533.

Bialystok, E. (1979). Explicit and implicit judgements of L2 grammaticality. *Language Learning, 29*(1), 81-103.

Bialystok, E. (1994). Representation and ways of knowing: Three issues in second language acquisition. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 549-569). San Diego, CA: Academic Press.

Bley-Vroman, R. (1990). The logical problem of foreign language learning. *Linguistic Analysis, 20*, 3-49.

Bley-Vroman, R. (2009). The evolving context of the fundamental difference hypothesis. *Studies in Second Language Acquisition, 31*(2), 175-198.

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. Tarone, S. Gass & A. Cohen (Eds.), *Research methodology in second language acquisition* (pp. 245-261).

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition, 33*(2), 247-271.

Brooks, P. J., Kempe, V., & Sionov, A. (2006). The role of learner and input variables in learning inflectional morphology. *Applied Psycholinguistics, 27*(2), 185-209.

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language, 61*(2), 171-190.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Buck, G. (1992). Listening Comprehension: Construct Validity and Trait Characteristics. *Language Learning, 42*(3), 313-357.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin, 56*(2), 81-105.

Carroll, J. B. (1981). Language aptitude measures in streaming, matching students with methods, and diagnosis of learning problems. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83-118). Rowley, MA: Newbury House.

Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge: UK: Cambridge University Press.

Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test: MLAT*. New York: : Psychological Corporation.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language, 47*(1), 30-49.

Chrabaszcz, A., & Jiang, N. (2014). The role of the native language in the use of the English nongeneric definite article by L2 learners: A cross-linguistic comparison. *Second Language Research, 30*(3), 351-379.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington Pub.

De Graaff, R. (1997). The eXperanto experiment. *Studies in Second Language Acquisition, 19*(2), 249-276.

DeKeyser, R. M. (1995). Learning second language grammar rules. *Studies in second language acquisition, 17*(3), 379-410.

DeKeyser, R. M. (1997). Beyond Explicit Rule Learning. *Studies in Second Language Acquisition, 19*(2), 195-221.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*(4), 499-534.

DeKeyser, R. M. (2003). Implicit and Explicit Learning. In C. J. Doughty & H. M. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 312-348). Oxford: Blackwell Publishers.

DeKeyser, R. M. (2007a). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. New York, NY: Cambridge Universtiy Press.

DeKeyser, R. M. (2007b). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 97–114). Mahwah, NJ: Erlbaum.

DeKeyser, R. M. (2009). Cognitive-Psychological Processes in Second Language Learning. In H. M. Long & C. J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 119-138). Oxford: Wiley-Blackwell.

DeKeyser, R. M. (2012). Interactions Between Individual Differences, Treatments, and Structures in SLA. *Language Learning, 62*(s2), 189-200.

DeKeyser, R. M. (in press). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (2nd ed.). Mahwah, NJ: Erlbaum.

DeKeyser, R. M., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics, 31*(3), 413-438.

DeKeyser, R. M., & Koeth, J. (2010). Cognitive Aptitudes for Second Language Learning. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (Vol. 2, pp. 395-406). New York: Routledge.

Destrebecqz, A., & Cleeremans, A. (2001). Can sequence learning be implicit? New evidence with the process dissociation procedure. *Psychonomic Bulletin & Review, 8*(2), 343-350.

Dienes, Z. (2004). Assumptions of subjective measures of unconscious mental states: Higher order thoughts and bias. *Journal of Consciousness Studies, 11*(9), 25-45.

Dienes, Z. (2007). Subjective measures of unconscious knowledge. *Progress in brain research, 168*, 49-269.

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological research, 69*(5-6), 338-351.

Dörnyei, Z. (2009). *The psychology of second language acquisition*: Oxford University Press Oxford.

Doughty, C. J., Campbell, S. G., Mislevy, M. A., Bunting, M. F., Bowles, A. R., & Koeth, J. T. (2010). *Predicting near-native ability: The factor structure and reliability of Hi-LAB.* Paper presented at the Selected Proceedings of the 2008 Second Language Research Forum: Exploring SLA Perspectives, Positions, and Practices.

Dussias, P. E., Valdés Kroff, J. R., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). When Gender and Looking Go Hand in Hand. *Studies in second language acquisition, 35*(2), 353-387.

Ellert, M. (2013). Resolving ambiguous pronouns in a second language: A visual-world eye-tracking study with Dutch learners of German. *International Review of Applied Linguistics in Language Teaching, 51*(2), 171-197.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(2), 143-188.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition, 27*(2), 305-352.

Ellis, N. C. (in press). Implicit AND Explicit Language Learning: Their dynamic interface and complexity. In P. Rebuschat (Ed.), *Investigating Implicit and Explicit Language Learning*: Routledge.

Ellis, R. (1990). *Instructed second language acquisition: Learning in the classroom*. Oxford: UK: Basil Blackwell.

Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language Learning, 54*(2), 227-275.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition, 27*(2), 141-172.

Ellis, R. (2006). Modelling Learning Difficulty and Second Language Proficiency: The Differential Contributions of Implicit and Explicit Knowledge. *Applied Linguistics, 27*(3), 431-463.

Ellis, R. (2008). *The Study of Second Language Acquisition* (2nd ed.). Oxford: Oxford University Press.

Ellis, R. (2009). Implicit and explicit learning, knowledge and instruction. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.), *Implicit and explicit*

knowledge in second language learning, testing and teaching (pp. 3-25). Tonawanda, NY: Multilingual Matters.

Ellis, R. (2009). Measuring implicit and explicit knolwedge of a second language. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 31-64). Tonawanda, NY: Multilingual Matters.

Ellis, R., & Loewen, S. (2007). Confirming the Operational Defnitions of Explicit and Implicit Knowledge in Ellis (2005): Responding to Isemonger. *Studies in second language acquisition, 29*(1), 119-126.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current directions in psychological science, 11*(1), 19-23.

Ercetin, G., & Alptekin, C. (2013). The explicit/implicit knowledge distinction and working memory: Implications for second-language reading comprehension. *Applied Psycholinguistics, 34*(4), 727-753.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics, 27*(3), 464-491.

Farmer, T. A., Anderson, S. E., & Spivey, M. J. (2007). Gradiency and visual context in syntactic garden-paths. *Journal of memory and language, 57*(4), 570-595.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: implications for group differences in response latency. *Psychological bulletin, 125*(6), 777-799.

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of memory and language*, *69*(3), 165-182.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language, 41*(1), 78-104.

Foote, R. (2011). Integrated knowledge of agreement in early and late English–Spanish bilinguals. *Applied Psycholinguistics, 32*(1), 187-220.

Foucart, A., & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language, 66*(1), 226-248.

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226-241.

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology, 2*, 1-6.

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics, 29*(3), 463-487.

Granena, G. (2010). *Elicited Imitation as a measure of implicit second language knowledge and working memory*. Unpublished Qualifying Paper, University of Maryland, College Park.

Granena, G. (2012). *Age differences and cognitive aptitudes for implicit and explicit learning in ultimate second language attainment.* Unpublished doctoral dissertation, University of Maryland, College Park, MD.

Granena, G. (2013a). Cognitive aptitudes for L2 learning and the LLAMA Language Aptitude Test. In G. Granena, & Long, M. H. (Ed.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 105-130). Philadelphia: PA: John Benjamins.

Granena, G. (2013b). Individual Differences in Sequence Learning Ability and Second Language Acquisition in Early Childhood and Adulthood. *Language Learning*, *63*(4), 665–703.

Granena, G. (2013c). Reexamining the robustness of aptitude in second language acquisition. In G. Granena, & Long, M. H. (Ed.), *Sensitive periods, language aptitude, and ultimate L2 attainment.* (pp. 179-204). Philadelphia: PA: John Benjamins.

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.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research*, *28*(2), 191-215.

Gutiérrez, X. (2013). The Construct Validity of Grammaticality Judgment Tests as Measures of Implicit and Explicit Knowledge. *Studies in second language acquisition, 35*(3), 423-449.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis with readings*. Englewood Cliffs, NJ: Prentice-Hall College.

Han, Y., & Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research, 2*(1), 1-23.

Han, Z., & Finneran, R. (2014). Re-engaging the interface debate: strong, weak, none, or all? *International Journal of Applied Linguistics, 24*(3), 370-389.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In S. d. T. R. Cudeck, & D. Sörbom (Ed.), *Structural Equation Modeling:*

Present and Future — A Festschrift in honor of Karl Jöreskog (pp. 195-216). Lincolnwood, IL: Scientific Software International, Inc.

Harley, B., & Hart, D. (1997). Language aptitude and second language proficiency in classroom learners of different starting ages. *Studies in Second Language Acquisition, 19*(03), 379-400.

Harley, B., & Hart, D. (2002). Age, aptitude and second language learning on a bilingual exchange. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 301-330). Amsterdam: John Benjamins.

Hasuike, I. (2004). Investigation on the Overuse of the Particle "Ni" for Location: Analysis of Particle Choice Strategies by Intermediate Chinese Speakers. [Basho wo arawasu kakujoshi 'ni' no kajou shiyou ni kansuru ichikousatsu: Chuukyuu reberu no chuugokuowasha no joshi sentaku sutoratejii bunseki]. *Nihongo Kyouiku*(122), 52-61.

Hasuike, I. (2012). Effects of L1 on the Choice between the Japanese Location Expressions Ni and De: Analysis of the Particle-Choice Test. [日本語の空間表現「に」と「で」の選択にみられる母語の影響ー助詞選択テストの結果分析ー]. *Kotoba to bunka, 13*, 59-76.

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research, 29*(1), 33-56.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: concepts, issues, and applications* (pp. 158-176). Thousand Oaks: Sage Publications.

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive psychology, 58*(3), 376-415.

Huang, Y. T., & Snedeker, J. (2013). The use of lexical and referential cues in children's online interpretation of adjectives. *Developmental Psychology, 49*(6), 1090-1102.

Huettig, F., Chen, J., Bowerman, M., & Majid, A. (2010). Do language-specific categories shape conceptual processing? Mandarin classifier distinctions influence eye gaze behavior, but only during linguistic processing. *Journal of Cognition and Culture, 10*(1-2), 39-58.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica, 137*(2), 151-171.

Hulstijn, J. H. (2002). Towards a unified account of the representation, processing and acquisition of second language knowledge. *Second Language Research, 18*(3), 193-223.

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning. *Studies in Second Language Acquisition, 27*(2), 129-140.

Hulstijn, J. H. (2007). Psycholinguistic perspectives on language and its acquisition. In J. Cummins & C. Davison (Eds.), *International Handbook of English Language Teaching* (pp. 701-713). Norwell, MA: Springer.

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.

Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in SLA. In C. Doughty & M. Long (Eds.), *Handbook of second language acquisition* (pp. 539-599). Oxford: Blackwell.

Inagaki, S. (2009). Acquisition of Tameni and Youni Indicating Purpose by Advanced L2 Japanese speakers with L1 Chinese. [Chuugokugo o bogo to suru joukyuu nihongogakushusha ni yoru mokuteki o arawasu 'tameni' to 'youni' no shutoku]. *Nihongo Kyouiku*(142), 91-101.

Iori, I., Takanashi, S., Nakanishi, K., & Yamada, T. (2000). *Japanese Grammar Handbook for Teaching Beginners' Level*. Tokyo: Three A Network.

Isemonger, I. M. (2007). Operational definitions of explicit and implicit knowledge: Response to R. Ellis (2005) and some recommendations for future research in this area. *Studies in second language acquisition, 29*(1), 101-118.

Iwasaki, S. (2002). *Japanese*. Philadelphia, PA: John Benjamins.

Jacobsen, W. M. (1992). *The transitive structure of events in Japanese*. Tokyo: Kuroshio.

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

January, D., Trueswell, J. C., & Thompson-Schill, S. L. (2009). Co-localization of Stroop and syntactic ambiguity resolution in Broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience, 21*(12), 2434-2444.

Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review/La Revue canadienne des langues vivantes, 64*(1), 215-238.

Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics, 25*(4), 603-634.

Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language learning. *Language Learning, 57*(1), 1-33.

Jiang, N. (2011). *Conducting Reaction Time Research in Second Language Studies*. New York, NY: Routledge.

Jiang, N., Hu, G., Lukyanchenko, A., & Cao, Y. (2010, October 14-17, 2010). *Insensitivity to morphological errors in L2: Evidence from word monitoring.* Paper presented at the SLRF 2010, College Park, MD.

Jiang, N., Novokshanova, E., Masuda, K., & Wang, X. (2011). Morphological Congruency and the Acquisition of L2 Morphemes. *Language Learning, 61*(3), 940–967.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive psychology, 21*(1), 60-99.

Jöreskog, K., & Sörbom, D. (2013). LISREL 9.1 for Windows. Skokie, IL: Scientific Software International: Inc.

Juffs, A. (1996). Semantics-syntax correspondences in second language acquisition. *Second Language Research, 12*(2), 177-221.

Juffs, A., & Harrington, M. (1995). Parsing effects in second language sentence processing. *Studies in second language acquisition, 17*(04), 483-516.

Kaiser, E., Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition, 112*(1), 55-80.

Kaiser, E., & Trueswell, J. C. (2008). Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes*, *23*(5), 709-748.

Kamide, Y., Altmann, G., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133-156.

Kamide, Y., Scheepers, C., & Altmann, G. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of psycholinguistic research*, *32*(1), 37-55.

Kanno, K., Hasegawa, T., Ikeda, K., & Ito, Y. (2005). Linguistic Profiles of Heritage Bilingual Learners of Japanese *Proceedings of the 4th International Symposium on Bilingualism* (pp. 1139-1151).

Kapnoula, E. C., Packard, S., Gupta, P., & McMurray, B. (2015). Immediate lexical integration of novel word forms. *Cognition*, *134*(0), 85-99. doi: http://dx.doi.org/10.1016/j.cognition.2014.09.007

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*(3), 321-340.

Keating, G. D. (2009). Sensitivity to Violations of Gender Agreement in Native and Nonnative Spanish: An Eye‐Movement Investigation. *Language Learning*, *59*(3), 503-535.

Klein, C., & Fischer, B. (2005). Instrumental and test–retest reliability of saccadic measures. *Biological psychology*, *68*(3), 201-213.

Kobayashi, N., Ford-Niwa, J., & Yamoto, H. (1996). SPOT: A new testing method of Japanese language proficiency. [Nihongo nouryoku no atarashii sokuteihou: SPOT]. *Japanese-Language Education around the Globe, 6*, 201-218.

Kotz, S. A. (2009). A critical review of ERP and fMRI evidence on L2 syntactic processing. *Brain and Language, 109*(2), 68-74.

Krashen, S. D. (1981). Second language acquisition and Second Language Learning. Oxford: Pergamon Press.

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon.

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

Krashen, S. D. (1994). The input hypothesis and its rivals. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 45-77). San Diego, CA: Academic Press.

Krashen, S. D., Long, H. M., & Scarcella, R. C. (1979). Age, rate and eventual attainment in second language acquisition. *Tesol Quarterly, 13*(4), 573-582.

Leung, J. H. C., & Williams, J. N. (2011). The Implicit Learning of Mappings between Forms and Contextually Derived Meanings. *Studies in Second Language Acquisition, 33*(1), 33-55.

Leung, J. H. C., & Williams, J. N. (2012). Constraints on Implicit Learning of Grammatical Form‐Meaning Connections. *Language Learning, 62*(2), 634-662.

Leung, J. H. C., & Williams, J. N. (2014). Crosslinguistic Differences in Implicit Language Learning. *Studies in second language acquisition, 36*(4), 733-755.

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science, 18*(3), 193-198.

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language, 63*(4), 447-464.

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

Lin, C. (2002). *Acquisition of Transitive/Intransitive Verb Pairs by Taiwanese Japanese L2 leaners.* Ochanomizu Women's University.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., . . . Doughty, C. J. (2013). Hi-LAB: A New Measure of Aptitude for High-Level Language Proficiency. *Language Learning, 63*(3), 530-566.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2013). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21*(4), 861-883.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24*(4), 489-515.

Loewen, S. (2009). Grammaticality judgment tests and the measurement of implicit and explicit L2 knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 94-112). Tonawanda, NY: Multilingual Matters.

Long, M. H. (2007). *Problems in SLA*. U.S.A.: Lawrence Erlbaum Associates.

Long, M. H. (2015). *Second Language Acquisition and Task-Based Language Teaching*. Oxford: Wiley-Blackwell.

Lukyanchenko, A. (2011). *The role of L1 transfer and the Explicit/Implicit interface in the acquisiiton of the English definite article.* Unpublished Qualifying Paper, University of Maryland, College Park.

Lyons, C. (1999). *Definiteness*. Cambridge, UK: Cambridge University Press.

Lyster, R., & Sato, M. (2013). Skill Acquisition Theory and the Role of Practice in L2 Development. In M. García Mayo, Gutierrez-Mangado, J., & Martínez Adrián, M. (Ed.), *Contemporary Approaches to Second Language Acquisition* (pp. 71-92). Amsterdam: John Benjamins.

Maeda, N. (2006). *Meaning and Usage of the 'youni Clause in Japanese*. Tokyo, Japan: Kasama Shoin.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8*(1), 1-71.

Martin, K. I., & Ellis, N. C. (2012). The Roles of Phonological Short-Term Memory and Working Memory in L2 Grammar and Vocabulary Learning. *Studies in Second Language Acquisition, 34*(3), 79-413.

Matin, E., Shao, K., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics, 53*(4), 372-380.

McLaughlin, B. (1987). *Theories of second-language learning*. London: Routledge.

Meara, P. M. (2005). LLAMA language aptitude tests. *Swansea: Lognostics*.

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy & L. E. Bourne Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339-364).

Morgan-Short, K. (2014). Electrophysiological Approaches to Understanding Second Language Acquisition: A Field Reaching its Potential. *Annual Review of Applied Linguistics, 34*, 15-36.

Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Working Papers on Bilingualism, 2*.

Nakaishi, Y. (2005). *Second Language Acquisition of Paried Intransitive/Transitive Verb: Unequality Usaeg of Verb Pairs.* Unpublished doctoral dissertation, Graduate School of Hiroshima University.

Nemeth, D., Janacsek, K., & Fiser, J. (2013). Age-dependent and coordinated shift in performance between implicit and explicit skill learning. *Frontiers in computational neuroscience, 7*, 1-13.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*(3), 417-528.

Novick, J. M., Thompson-Schill, S. L., & Trueswell, J. C. (2008). Putting lexical constraints in context into the visual-world paradigm. *Cognition, 107*(3), 850-903.

Paradis, M. (1994). Neurolinguistic aspects of implicit and explicit memory: implications for bilingualism. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 393-419). San Diego, CA: Academic Press.

Paradis, M. (2004). *A neurolinguistic theory of bilingualism* (Vol. 18): John Benjamins Publishing Company.

Paradis, M. (2009). *Declarative and procedural determinants of second languages*. Philadelphia, PA: John Benjamins Publishing Company.

Pienemann, M. (1998). *Language processing and second language development: Processability theory*. Amsterdam: John Benjamins Publishing.

Pienemann, M. (2005). *Cross-linguistic aspects of Processability Theory*. Amsterdam: John Benjamins Publishing Company.

Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery (PLAB)*. Washington, DC: Second Language Testing Incorporated.

Reber, A. S., Walkenfeld, F. F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(5), 888-896.

Rebuschat, P. (2013). Measuring Implicit and Explicit Knowledge in Second Language Research. *Language Learning, 63*(3), 595–626.

Rebuschat, P., Hamrick, P., Sachs, R., Riestenberg, K., & Ziegler, N. (2013). Implicit and explicit knowledge of form-meaning connections: Evidence from subjective measures of awareness. In J. Bergsleithner, S. Frota & J. K. Yoshioka (Eds.), *Noticing: L2 studies and essays in honor of Dick Schmidt* (pp. 249-270). Honolulu, HI: University of Hawai'i at Manoa, National Foreign Language Resource Center.

Rebuschat, P., & Williams, J. N. (2011). Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics, 33*(4), 829-856.

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(3), 585-594.

Roberts, L., & Liszka, S. A. (2013). Processing tense/aspect-agreement violations on-line in the second language: A self-paced reading study with French and German L2 learners of English. *Second Language Research, 29*(4), 413-439.

Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in second language acquisition, 19*(2), 223-247.

Robinson, P. (2005). Cognitive abilities, chunk-strength and frequency effects during implicit Artificial Grammar, and incidental second language learning: Replications of Reber, Walkenfeld and Hernstadt (1991) and Knowlton and Squire (1996) and their relevance to SLA. *Studies in Second Language Acquisition, 27*(2), 235-268.

Rodgers, D. M. (2011). The automatization of verbal morphology in instructed second language acquisition. *IRAL-International Review of Applied Linguistics in Language Teaching, 49*(4), 295-319.

Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2003). Assignment of reference to reflexives and pronouns in picture noun phrases: evidence from eye movements. *Cognition, 89*(1), B1-B13.

Sarandi, H. (2015). Reexamining elicited imitation as a measure of implicit grammatical knowledge and beyond…? *Language Testing*. doi: 10.1177/0265532214564504

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355-390.

Schmidt, R. (1990). The Role of Consciousness in Second Language Learning. *Applied Linguistics, 11*(2), 129-158.

Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics, 13*(1), 206-226.

Schmidt, R. (1994a). Deconstructing consciousness in search of useful definitions for applied linguistics. *Consciousness in second language learning, 11*, 237-326.

Schmidt, R. (1994b). Implicit learning and the cognitive unconscious: Of artificial grammars and SLA. In N.Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 165-209).

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). New York, NY: Cambridge University Press.

Sedivy, J. C. (2010). Using eyetracking in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental Methods in Language Acquisition Research* (pp. 115-138). Philadelphia: PA: John Benjamins Publishing Company.

Shanks, D. R., & Johnstone, T. (1999). Evaluating the relationship between explicit and implicit knowledge in a sequential reaction time task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(6), 1435-1451.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Spada, N., & Tomita, Y. (2010). Interactions Between Type of Instruction and Type of Language Feature: A Meta-Analysis. *Language Learning, 60*(2), 263-308.

Suzuki, Y., & DeKeyser, R. M. (in press). Comparing elicited imitation and word monitoring as measures of implicit knowledge. [Unpublished Qualifying Paper]. *Language Learning*.

Tanenhaus, M. K., Chambers, C. C., & Hanna, J. E. (2004). Referential domains in spoken language comprehension: Using eye movements to bridge the product and action

traditions. *The interface of language, vision, and action: Eye movements and the visual world*, 279-317.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632-1634.

Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia, 56*, 289-301.

Tomita, Y., Suzuki, W., & Jessop, L. (2009). Elicited imitation: Toward valid procedures to measure implicit second language grammatical knowledge. *TESOL Quarterly, 43*(2), 345-350.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications* (Vol. 3): Sage.

Trenkic, D., Mirkovic, J., & Altmann, G. T. M. (2014). Real-time grammar processing by native and non-native speakers: Constructions unique to the second language. *Bilingualism: Language and Cognition, 17*(2), 237-257.

Trofimovich, P., Ammar, A., & Gatbonton, E. (2007). How effective are recasts? The role of attention, memory, and analytical ability. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 171-195). Oxford: New York: Oxford University Press.

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition, 73*(2), 89-134.

Vinther, T. (2002). Elicited imitation: a brief overview. *International Journal of Applied Linguistics, 12*(1), 54-73.

Wen, Z., & Skehan, P. (2011). A new perspective on foreign language aptitude research: building and supporting a case for "working memory as language aptitude". *A Journal of English Language, Literatures in English and Cultural Studies, 60*, 15-44.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*(1), 1-26.

Williams, J. N. (1999). Memory, attention, and inductive learning. *Studies in Second Language Acquisition, 21*(1), 1-48.

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition, 27*(2), 269-304.

Williams, J. N. (2009). Implicit learning in second language acquisition. *The new handbook of second language acquisition*, 319-353.

Williams, J. N., & Lovatt, P. (2003). Phonological memory and rule learning. *Language Learning, 53*(1), 67-121.

Woltz, D. J. (2003). Implicit cognitive processes as aptitudes for learning. *Educational Psychologist, 38*(2), 95-104.

Zhang, R. (2014). Measuring University-Level L2 Learners' Implicit and Explicit Linguistic Knowledge. *Studies in second language acquisition, First View*, 1-30.