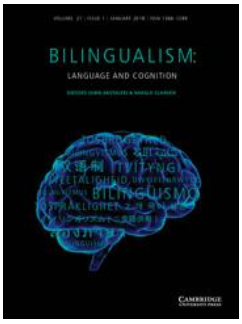
Running head: Automatization in L2 sentence processing

Automatization in second language sentence processing: Relationship between elicited imitation and maze tasks



Yuichi Suzuki and Midori Sunada

Author Notes: Yuichi Suzuki, Faculty of Foreign Languages, Kanagawa University; Midori Sunada, Graduate School of Education, Tokyo Gakugei University.

Acknowledgements: We would like to express our gratitude to Professors Yoshiki Takayama, Misato Usukura, and Tetsuo Baba for their generous cooperation in data collection. We are very grateful to Ms. Kanno for her assistance in data coding. We wish to thank the Associate Editor, Prof. Ludovica Serratrice, and the three anonymous reviewers for providing insightful and constructive feedback.

Correspondence concerning this article should be addressed to Yuichi Suzuki, Faculty of Foreign Languages, Kanagawa University, 3-27-1, Rokkakubashi, Kanagawa-ku, Yokohama-shi, Kanagawa, 221-8686, JAPAN. Email Address: szky819@kanagawa-u.ac.jp

Keywords: Second language acquisition, Automatization, Coefficient of variance, Elicited imitation, Maze task

Abstract

The present study investigates the automatization of second language (L2) sentence processing. It compares the extent to which a mere speedup (faster execution) and restructuring (more stable execution) of sentence processing contribute to L2 oral performance. The maze task is used to measure the speed (reaction time, RT) and processing stability (coefficient of variance, CV) of sentence processing. The elicited imitation (EI) task measures L2 oral proficiency (repetition accuracy and accuracy in plural and third person *s*). These tasks were performed by 110 English-as-a-foreign-language learners with Japanese as their L1. The results show that only RT, not CV, significantly predicts L2 oral proficiency. Even though a subgroup of learners, who previously stayed in an English-speaking country, demonstrated some indications of automatization, RT was a better predictor of L2 oral proficiency than CV, irrespective of immersion experience. These findings suggest that CV has little practical value in predicting L2 oral proficiency.

Introduction

Smooth engagement in L2 communication requires complex, coordinated lower-level sub-skills, such as lexical retrieval, grammatical parsing, and articulating sounds (Kormos, 2006; Levelt, 1989, 1999). Automaticity of these lower-level sub-skills plays a critical role in supporting fluent L2 use (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2013). Skill acquisition theory postulates that L2 learners gradually automatize these skills, progressing from controlled to more automatic processes (Anderson, 2015; DeKeyser, 2015; McLaughlin, 1987). These automatization processes have been documented for L2 learning (e.g., DeKeyser, 1997), which follows developmental paths similar to those of other cognitive skills, such as mathematical calculation, playing a sport, and driving a car, among others. However, as Lim and Godfroid (2015) noted:

> Many important questions have remained unanswered as to how L2 automaticity develops over the course of language learning, how long it takes for L2 learners to reach the fully automatized phase, how the development of automaticity varies depending on linguistic features, or how automaticity can be validly measured. (p. 1248)

The current study aims to address the assessment issues of automaticity in L2 sentence processing. A body of research has attempted to measure automaticity at the word level (i.e., lexical access), using Reaction Time (RT) tasks, such as lexical decision tasks and semantic classification tasks (Akamatsu, 2008; Harrington, 2006; N.S. Segalowitz & Freed, 2004; N. S. Segalowitz & S.J. Segalowitz, 1993; N. S. Segalowitz, Watson, & S.J. Segalowitz, 1995; S. J. Segalowitz, N.S. Segalowitz, & Wood, 1998).

In contrast, very few studies have attempted to assess automaticity at the sentence level (DeKeyser, 1997; Hulstijn, Van Gelderen, & Schoonen, 2009; Lim & Godfroid, 2015; Rodgers, 2011). By extending this line of recent work on the assessment of automaticity in L2 sentence processing, the present study measures automaticity in sentence (syntactic) processing[1] using a computerized RT task called a maze task. The current cross-sectional study investigates the extent to which the sentence processing efficiency, as measured by this maze task, can predict L2 oral proficiency (as measured by the Elicited Imitation [EI] task).

Automatization and CV in L2 Learning

Automaticity is defined as a fast, ballistic, effortless, and unconscious process (N. S. Segalowitz, 2003). A seminal study by N. S. Segalowitz and S. J. Segalowitz (1993) proposed a distinction between fast and automatic L2 processing. This distinction can also be characterized as a mere speedup and stable processing (i.e., automatization). The authors claimed that a faster processing speed or *quantitative* change is necessary, but not sufficient for achieving genuine automatic processing. In other words, the automatization of language processing also entails a *qualitative* shift in processing. For instance, in the domain of English syntax, several processes may first mediate the assembly of words to formulate a sentence, such as L1 translation, re-ordering of L1 word order to one appropriate for L2, and use of declarative knowledge (e.g., knowledge about grammatical rules). While a mere speedup indicates a faster execution of these processes, automaticity indicates a restructuring of syntactic processing; for example, by means of bypassing (some of) these processes. Restructuring (i.e., more stable processing) can be indexed according to the coefficient of variance (CV), which is computed by dividing the mean standard deviation (SD) of RT for each individual by his/her mean RT (N. S. Segalowitz & S. J.

Segalowitz, 1993). Extant studies have proposed the variability scores expressed by CV as a useful index for automatization.

The distinction between a mere speedup and genuine automatization bears potential practical implications for examining and assessing L2 learning processing. Yet, very little is presently known about the effect of qualitative shifts in automatic sentence processing on L2 skills, such as speaking skills. On the one hand, CV, which is postulated to measure qualitatively different processing, may reveal genuine automatization and serve as a useful index for predicting speaking abilities. On the other hand, some SLA researchers question the usefulness of CV because the mathematical distinction between RT and CV is very subtle (Hulstijn et al., 2009).

The current study re-examines the validity of CV as an index of automaticity in L2 sentence processing. To meet this purpose, it tested three conditions that are needed to be met for automatization: (1) faster RT, (2) a positive RT–CV correlation, and (3) smaller CV. Subsequently, we elaborate why (2), a positive RT–CV correlation, is postulated to indicate automatization. The mean of both RT and SD usually decreases as the execution of the task becomes faster. If the processing becomes "just faster," then CV does not substantially change, as RT and SD decrease at a similar rate. In other words, a combination of faster RT and unchanged CV can never produce a positive correlation. However, if the processing becomes more "automatic," it leads to a disproportional reduction of SD relative to RT, resulting in smaller CV and producing a positive correlation between RT and CV (see N. S. Segalowitz & S. J. Segalowitz, 1993 for detailed explanations with numerical examples). A positive correlation indicates that some processing components are bypassed or eliminated, and this criterion can be used as a litmus test for automatization (Hulstijn et al., 2009).

Previous Research on Automatization in L2 Sentence Processing

Compared to research on lexical access, relatively fewer studies have attempted to measure automaticity at the sentence level (DeKeyser, 1997; Hulstijn et al., 2009; Lim & Godfroid, 2015; Rodgers, 2011). The most relevant contributions in this field stem from the pioneering experimental work of Hulstijin et al. (2009) and its subsequent conceptual replication by Lim and Godfroid (2015). In both studies, the researchers use a common computerized RT task, called a sentence construction (production) task, to assess the automatization of L2 sentence processing. In this sentence construction task, after the first fragment of a sentence (e.g., What does that) was presented to the participants; they were prompted to choose the word that grammatically continues the first phrase as quickly as possible, by selecting one of the two options presented (e.g., A. *she,* B. *mean*). The task's objective is to measure how rapidly learners can build sentence structures by selecting the correct option.

By means of a two-year longitudinal study, Hulstijin et al. (2009) investigated to what extent high school students, with Dutch as L1 and English as L2, develop L2 automatic processing. The same cohort of about 200 learners was subsequently tested by the sentence construction task through Grades 8, 9, and 10. The results show that, across those two years, (a) RT significantly decreased, (b) the CV–RT correlation was negative and ranged from none to small ($-.004 < r < -.342$), and (c) CV remained unchanged. Overall, these findings did not support the usefulness of CV as a measure of automaticity at the sentence level, which contrasts the findings at the word level (e.g., N. S. Segalowitz & S. J. Segalowitz, 1993; for detailed reviews on lexical access, see Hulstijn et al., 2009).

Lim and Godfroid (2015) performed a conceptual replication of the experiment conducted by Hulstijn et al. (2009) to further investigate whether or not CV can index automatization in L2 sentence processing. They recruited 40 English L2 learners with Korean as L1, most of who were undergraduate or graduate students at a university in the United States. Unlike Hulstijn et al. (2009), Lim and Godfroid (2015) employed a cross-sectional design and assessed learners' proficiency levels through a diagnostic test (i.e., the vocabulary and grammar sections of the DIALANG). Based on the DIALANG scores, the learners were classified as advanced (corresponding to C1 or C2 in the Common European Framework, Council of Europe, 2001) or as intermediate (corresponding to B1 or B2 in the same). The results of the sentence construction task show that (a) RT was significantly faster for advanced learners than for intermediate learners, (b) the CV–RT correlation was significant in the group that comprised advanced learners (r = .713), but not in the intermediate group (r = .189), and (c) CV was significantly lower for the advanced group relative to the intermediate group. Contrary to the findings reported by Hulstijn et al. (2009), these results support N. S. Segalowitz and S. J. Segalowtiz's (1993) hypothesis that CV can capture L2 automatization.

When interpreting these conclusions, it is essential to note a number of differences between Hulstijn et al.'s (2009) study and its conceptual replication by Lim and Godfroid (2015). We will delineate three primary factors here to guide the present study (see Lim & Godfroid, 2015, for a more extensive discussion). First, automatization in syntactic processing can be obscured by a positive transfer from a L1 that is typologically similar to L2 (see Koda, 2007, for review). Therefore, Dutch learners are more likely to benefit from a positive transfer from their L1 knowledge of orthography and syntactic processing to L2 English sentence processing. In contrast, Korean learners have a typologically different L1 in terms of both orthography and syntactic structures, and therefore face greater challenges in automatizing syntactic processing. Consequently, automatization might have a more profound effect on L2 proficiency for Korean learners than it has for Dutch learners. Secondly, the skill learning stage or proficiency may be more advanced for the ESL learners that took part in Lim and Godfroid's (2015) study, compared to the Dutch high school students that participated in Hulstijn et al.'s (2009) pioneering work.[2] Thirdly, the majority of the learners that formed the sample in Lim and Godfroid's (2015) study had been immersed in an English-speaking environment. An immersion context is likely to facilitate automatization because it provides ample opportunities for extensive practice (DeKeyser, 2007).

The present study builds on the aforementioned studies. It adopts a cross-sectional design and its participants are English L2 learners with a typologically different L1 background (i.e., Japanese). L2 proficiency is measured by means of an objective L2 oral assessment tool (EI task), and the study also examines the role of immersion experience.

The Maze Task as a Measure of Automatization in L2 Sentence Processing

Following the rationale of the sentence construction task, the present study developed an alternative, a psycholinguistic task called the maze task (Forster, Guerrera, & Elliot, 2009; Witzel, Witzel, & Forster, 2012), which is a useful measure for the automatization of sentence processing. In this maze task, participants have to construct an entire sentence by choosing from two options. As shown in Figure 1, participants are presented with two words as options side by side and are asked to choose the word that correctly continues the sentence. Similar to the sentence construction task, one of the options is correct, while the other option is ungrammatical

and incorrect. The maze task thus requires learners to constantly predict the next word and immediately integrate the previous word. Unlike the sentence construction task, the maze task can assess the real-time incremental processing of the whole sentence, rather than only a fragment of the sentence.

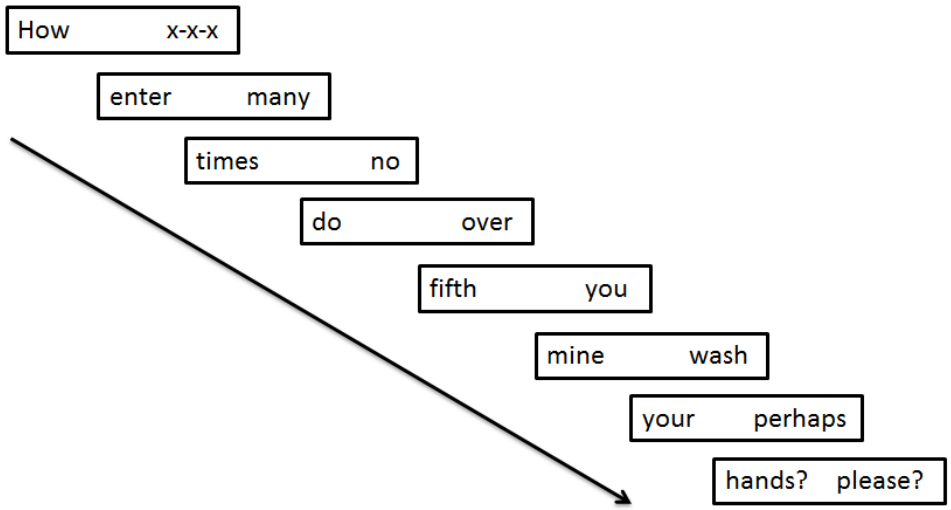


Figure 1. A sample display of the maze task.

As the maze task has been developed as part of L1 psycholinguistic research, very few studies have utilized it to assess L2 processing (Enkin, 2012; McBride, 2011). For instance, Enkin and Forster (2014) used the maze task both as a training and outcome assessment measure. Beginner-level Spanish learners were trained in Spanish sentence structures that differ from those present in English. After three training sessions, the learners were able to perform the maze task faster. To the best of our knowledge, CVs have not been previously computed for the maze task. Nonetheless, the findings of pertinent studies are encouraging as they indicate that the maze task is sensitive enough to capture when more rapid sentence processing occurs in L2.

The Elicited Imitation Task as a Measure of L2 Speaking Proficiency and Automatized Grammatical Knowledge

An important question regarding the L2 sentence automatization process is to what extent speedup (RT) and restructuring (CV), as measured by the maze task, are related to linguistic skills, such as speaking skills. Faster sentence processing has been found to predict fluency in L2 speech (De Jong et al., 2013). However, to date, the extent to which genuine automatization, as indexed by CV, can predict L2 speaking proficiency has not been investigated. If restructuring or qualitative changes in automatization indeed take place in L2 grammar learning, as claimed by N. S. Segalowitz and others, they may influence speaking performance.

The present study compares the speed measures from the maze task with L2 oral proficiency as measured by the EI task. The EI task requires participants to listen to a sentence and subsequently repeat it as accurately as they can. It is primarily used to assess two aspects of L2 oral proficiency: (a) global oral L2 proficiency and (b) grammatical knowledge. Firstly, EI performance is often scored for repetition accuracy, and it has proven to be a reliable and useful measure for global oral L2 proficiency (Ortega, Iwashita, Norris, & Rabie, 2002; Tracy-Ventura, McManus, Norris, & Ortega, 2014; Wu & Ortega, 2013; Yan, Maeda, Lv, & Ginther, 2015).

Secondly, the EI task has also been used to assess grammatical knowledge for specific grammatical structures (Erlam, 2006; Spada, Shiu, & Tomita, 2015; Suzuki & DeKeyser, 2015). For this purpose, a target structure is embedded in a stimulus sentence, and the task assesses whether learners can repeat the target grammatical structure spontaneously under time pressure. It is assumed that participants draw on automatized grammatical knowledge by that time-pressured procedure irrespective of whether that grammatical knowledge is explicit or implicit (see Suzuki & DeKeyser, 2015, for further discussion).

The benefits of automatization in sentence processing extend by making more cognitive resources available (N. S. Segalowitz, 2010). Less or non-automatic processing overloads short-term memory while more automatic sentence processing frees up a certain amount of cognitive resources for L2 comprehension and production. In addition, L2 learners with more cognitive resources at their disposal may be able to attend to less salient linguistic features during the EI task. Hence, we explore the relation between automaticity in sentence processing and the accurate use of grammatical features by embedding a target grammatical feature in an EI stimulus sentence. The target grammatical features include the third person *s* and plural *s* because L2 English learners, whose L1 does not grammatically encode these, find them very hard to use spontaneously (Ellis, 2009). Automatization may play an important role in the accurate use of those morphological features.

The Current Study

The current study, which employs a cross-sectional design, administers the maze and EI tasks to EFL learners whose L1 is Japanese (which is typologically different from English). The aim is to examine to what extent RT and CV from the maze task can predict L2 speaking proficiency as measured by the EI task. Unlike previous studies, we use the EI task to measure multiple dimensions of L2 proficiency: (a) overall L2 oral proficiency and (b) grammatical knowledge of specific morphological features (third person *s* and plural *s*). We propose that CV, by capturing the qualitative changes in L2 syntactic processing, could be a useful predictor of oral L2 performance in the EI task. In other words, the elimination of unnecessary routines (e.g., re-ordering from L1 syntax to L2 syntax) might play a more crucial role in the improvement of L2 speaking performance than mere speedup, as speaking is cognitively highly demanding. Alternatively, CV might not be as informative as other measures, such as RT (Hulstijn et al., 2009). As a part of the study, we also examine the role of immersion experience in order to isolate the effect of context, which is conducive for automatization. The current study addresses the following research questions:

1. To what extent do RT and CV predict speaking skills (accuracy in repetition and use of morphological markings)?
2. Is there a positive correlation between RT and CV?
3. Does immersion experience influence the automatization process?

Methods

Participants

The study participants comprised 110 English L2 learners with Japanese as L1, who were recruited from four English classes at a national teacher's college in Japan. All participants have received at least six years of classroom instructions. All of them were taking courses related to

English education in order to obtain a teacher's license for English teaching, and their proficiency is considered relatively high for a regular group of Japanese university students. Four participants were excluded from the analysis due to technical problems, and an additional four participants were removed because they failed to follow the instructions of the maze task. This led to a final study sample of 102 participants, whose data was subjected to analysis. The mean age of the final sample was 20.18 (*SD* = 1.16). Approximately, a third of the participants reported that they had stayed in an English speaking country before or during their time at college (*n* = 33). The mean length of stay was 15 months (*SD* = 24.31). The current study did not recruit native speakers, as we expected that their EI performance would be near perfect, which would make it impossible to compare their EI scores with RT and CV from the maze task.

Instruments

Stimuli and target structures

The maze and EI tasks assessed sentence processing with four types of macro syntactic structures: declaratives, main wh-questions, relative clauses, and indirect questions. Table 1 presents the sample sentences. These four structures were chosen in order to assess the L2 processing speed of a variety of syntactic structures. Instead of focusing solely on one type of structure (e.g., declaratives), these four types of structures reflect a more representative sample of English sentences. The third person *s* was embedded in declarative and relative clause sentences, whereas the plural *s* was embedded in sentences with main wh-questions and indirect questions. An equal number of syntactic structures and morphological structures were created.

Table 1. *Syntactic structures and sample sentences used in the EI and maze tasks*.

Syntactic Structure	Morph. Feature	Sample Sentence	Num. of Items
Declarative (DEC)	3rd person	My brother always eats breakfast in the morning.	6
		(Simple)	6
		Your health becomes worse if you eat too much. (Complex)	
Question (QUE)	Plural	When did you learn to play so many songs on the piano?	12
Relative clause (REL)	3rd person	The lady knows the shop which is popular in Tokyo.	6
		(Subject REC)	6
		John's sister goes to a university which many people know. (Object REL)	
Cancel inversion	Plural	A lot of foreigners asked whether the train was going	12

(INV) south.

Two lists, each consisting of 48 sentences, were constructed for each task. The stimulus sentences were constructed carefully by keeping the number of words and syllables equal across both lists (see Appendix A). Similarly, in the EI task, the speech length and words (syllables) per minute in audio recording were almost identical for both lists. According to the results yielded by independent t-tests, there was no significant difference in variables between the two lists, $p > .1$. Furthermore, only words that were familiar to the participants were selected, in order to minimize the influence of lexical knowledge on the task performance. The lexical items were chosen from English textbooks for Japanese junior and high schools. The percentage of vocabulary coverage regarding the most frequent 2,000 words was also checked (Cobb, 2002). The first 1,000 most frequent words covered 90.73%, the second 1,000 most frequent words covered the next 4.69%, and the remaining 4.59% were outside the range. All verbs were within the most frequent 2,000 word range. All words outside the list of 2,000 words were either loan words that are in common use in Japanese (e.g., *classmates, hamburgers*) or familiar proper nouns (e.g., *Mary, John, Japan, London*). The coverage of the frequent words was also similar for both lists (89.39%, 5.10%, and 5.51% in List 1; 91.88%, 4.36%, and 3.76% in List 2). Appendix S1 presents all stimulus sentences as part of the online supplementary materials.

Maze task

As Figure 1 above shows, the first screen presented participants with two options. The left word was always the first word of the sentence, and the cross sign was presented on the right (e.g., *How* and *x-x-x*). The second screen asked them to quickly choose the correct word from the two words that would follow the previous word by pressing either the left (f) or the right (j) button (e.g., *enter* and *many*). If they gave a correct response, the next option of two words appeared on the screen. This process continued until the end of the sentence. When participants chose a wrong word, the trial automatically ended and the remaining options for that sentence were skipped. Therefore, when the participants chose a wrong answer, no response data were collected for the words that would have followed. Participants were told to respond as quickly and accurately as possible. Four practice items were presented before the actual test to familiarize participants with the task procedure. The duration of the maze task was approximately 10 to 15 minutes.

EI task

In the EI task, the participants were asked to (a) listen to a stimulus sentence through headphones, (b) say "three, two, one" in English when these numbers were presented on the computer screen in that respective order, and (c) repeat the sentence as exactly as they could. After listening to a sentence, participants were requested to count aloud three times in order to prevent the rote memorization of linguistic forms in short-term memory (Mackey & Gass, 2005; Suzuki & DeKeyser, 2015). This strategy ensured that participants had to reconstruct the sentence by using their long-term memory or linguistic knowledge.

The recording length of the stimuli was also calibrated to exceed the capacity of short-term memory. Since it is often possible to memorize words with lengths of 1.5–2.0 seconds (Baddeley, Thomson, & Buchanan, 1975; Cowan et al., 1992), all test items' recordings were longer than 2 seconds (see Appendix A for information about the stimulus sentences).

Time pressure was imposed during the repetition phase, whereby the time limit was set at twice the length of each recorded sentence (Jensen & Vinther, 2003). The time limits ranged between 4.68 and 9.06 seconds. At the end of the time allocated for repetition, a bell chime rang

and the trial automatically moved on to the next item. All the stimulus sentences were recorded by a female native speaker of American English who uttered them at a natural speed (see words per minute in Appendix A). The EI task took approximately 15 minutes to complete.

Procedures

Data collection was conducted in a large computer lab during regular class hours. The participants (22 to 30 students for each class) performed the two tasks individually. Both the EI and maze tasks were administered through DMDX software (Forster & Forster, 2003). For the EI task, the students put on headphones and were asked to hold an IC recorder close to their mouth for recording. Since there were four different classes, the order of the tasks and the lists were counter-balanced. More specifically, Class 1 first took the maze task (List 1) followed by the EI task (List 2), Class 2 took the maze task (List 2) followed by the EI task (List1), Class 3 took the EI task (List 1) followed by the maze task (List 2), and Class 4 took the EI task (List 2) followed by the maze task (List 1).

Coding and Data Analysis

Maze task

In the maze task, the responses to each word-item were collected and the data were analyzed at both the word and sentence level. For the word-level analysis, accuracy and RT were analyzed for each word-item (448 items and 456 items in List 1 and 2, respectively). RT was included in the analysis only when the response was correct. For the sentence-level analysis, a more stringent procedure was adopted, whereby a credit for accuracy was given only when *all* word-items for a given sentence were answered correctly (48 items). RT for each sentence-item was calculated by summing up RT for all word-items, but it was computed only when all responses were correct for a given sentence. If a participant made an error regarding any of the word-items in a given sentence, RT for that sentence was omitted from the total. Note that the first item of each sentence was excluded from the analysis, since it was already given to the participants (see Instruments section).

In addition, following the analysis procedures that were used in previous studies (Hulstijn et al., 2009; Lim & Godfroid, 2015), two types of analysis were conducted separately, namely, (1) raw data analysis and (2) cleaned data analysis. For the raw data analysis, only RTs of correct responses were analyzed and no additional data cleaning procedures were employed. The cleaned data analysis addressed missing (incorrect) and outlying responses by using multiple imputation (Rubin, 1977). In this study, participants who scored below 62.5% were excluded from the analysis to facilitate the assessment of the processing speed/variability with less influence from (declarative) linguistic knowledge. Secondly, items with fewer than 75% correct responses were excluded for the same reason. Thirdly, RTs for each item were inspected for outliers, and responses that were too fast or too slow were identified and categorized as missing. The low cutoff value was set at 300 ms,[3] and the high cutoff value was set as 3SD above the group mean for each item. Finally, all missing RTs (i.e., those pertaining to incorrect responses or outliers) were estimated by multiple imputation. Missing data were imputed within the maximum and minimum RT range for each item (Lim & Godfroid, 2015). Appendix B presents these data cleaning results.

EI task

The EI performance was coded in terms of (1) repetition accuracy and (2) use of targeted morphological structures. For the first point, the overall repetition accuracy was scored using a five-point scoring rubric that was developed and validated in previous EI studies (Ortega et al., 2002; Tracy-Ventura et al., 2014; Wu & Ortega, 2013):

4 = Perfect repetition
3 = Accurate content repetition with some (ungrammatical or grammatical) changes of form
2 = Changes in content or in form that affect meaning
1 = Repetition of half of the stimulus or less
0 = Silence, only one word repeated or unintelligible repetition

The primary trait for scoring focused on meaning or content conveyed in repetition. Perfect repetition was assigned a score of 4. As long as the original meaning was preserved, a sentence with some changes in form scored 3, regardless of the sentence's grammaticality. Here, substitutions with synonyms were also accepted (e.g., "a lot of" for "many"). A sentence involving substitutions or omissions that changed the original meaning received a score of 2. When a sentence retained only half the idea units or lexical items or less, it received a score of 1. No credit was given to silence or minimal repetition, which included one content word or only function words. Responses in which a participant failed to say the numbers ("three, two, one") aloud were also scored as incorrect (these accounted for only 0.57% of the total responses).

The second coding aim was to assess the accuracy of morphological markings (plural *s* and third person *s*). Based on the analysis protocol adopted in previous research (Erlam, 2006; Suzuki & DeKeyser, 2015), all sentences were categorized into three types of responses: (a) obligatory context created for a target morphological feature—supplied (correct), (b) obligatory context created—not supplied (incorrect), and (c) no obligatory context created (incorrect). The sample responses are presented below:

Sample response with plural *s*
Stimulus sentence: The boy wondered if he should take three classes at school.
(1a) The boy wondered if he should take ***three classes***. (Correct)
(1b) The boy wondered if he should take ***three class***. (Incorrect)
(1c) The boy wondered if he should take ***a class*** at school. (Incorrect)

Sample response with third person *s*
Stimulus sentence: My teacher often asks a lot of difficult questions in class.
(2a) My ***teacher*** often ***asks*** many difficult questions in class. (Correct)
(2b) My ***teacher*** often ***ask*** a lot of questions in class. (Incorrect)
(2c) My ***teachers*** often ***ask*** many questions in class. (Incorrect)

A response was scored as correct (1a or 2a) when the obligatory context for a target morphological feature was present and when the appropriate morphological form was supplied. Based on the second criterion, no credit was given for a response that lacked a target

morphological marker (plural *s* or third person *s*) (1b or 2b). For the third criterion, we deemed that no obligatory context was present when a response included *a* instead of the numeral in the stimulus (1c) or when the subject of the sentence was plural (2c). No credit was awarded for the third criterion.

Two raters independently coded the same subset (16%) of all responses. Both raters were Japanese native speakers with advanced English proficiency; an undergraduate and a graduate student who majored in Teaching English as a Foreign Language (TEFL). Their coding results were subsequently compared which revealed an inter-rater agreement of 96% for repetition accuracy and one of 100% for coding morphological structures (plural *s* and third person *s*). Disagreements were resolved through discussion, and the remainder of the data set was divided in half and separately scored by the two raters. The internal consistency indexed by Cronbach alpha was high for repetition accuracy (.96) and morphological accuracy (.88).

Comparison between the maze and the EI task

As stated above, three conditions were tested to examine the automatization process: (1) faster RT, (2) a positive RT–CV correlation, and (3) smaller CV. First, the maze task's RT was compared with all EI measures (i.e., repetition accuracy and the use of plural *s* and third person *s*). The correlations between RT and the aforementioned EI measures were expected to be negative, because faster RT should lead to higher accuracy in EI performance.[4] Secondly, RT and CV were correlated. Thirdly, the correlations between CV and EI measures were computed. In line with the first case of RT, the presence of a negative correlation was expected, because a smaller CV would lead to better EI performance. After inspecting the correlational patterns among RT, CV, and EI measures, multiple regression analyses were conducted to investigate to what extent RT and CV can predict EI scores. All analyses were first conducted using the data pertaining to the entire study sample, followed by separate analyses based on the participants' immersion experience.

Results

Table 2 presents the descriptive statistics for the EI and maze tasks. As the mean accuracy of the EI task was 1.6 (out of 4), it was somewhat difficult for the participants. However, the maze task accuracy is more relevant for the present investigation because Lim and Godfroid (2015) suggested that automatization requires a nearly perfect accuracy rate in order to isolate the improvement in RT. The percentage accuracy score of the maze task indicated an accuracy of only 87% and 77% at the word and sentence level, respectively. The data cleaning procedure would be expected to influence the results (Hulstijn et al., 2009; Lim & Godfroid, 2015).

Table 2. *Descriptive statistics for EI and maze tasks.*

Measures	Mean	SD	Min.	Max.
EI				
Repetition Accuracy	1.60	0.62	0.38	3.06
Morphological Accuracy	0.49	0.18	0.11	0.88
Third-Person-s Accuracy	0.43	0.19	0.04	0.88
Plural-s Accuracy	0.55	0.21	0.09	0.96
Maze (Word-level analysis)				
Accuracy	0.87	0.09	0.37	0.99
RT	1144	179	734	1829
SD	482	94	255	700
CV	0.42	0.05	0.26	0.50
Maze (Sentence-level analysis)				
Accuracy	0.77	0.14	0.29	0.96
RT	10742	1669	6913	17608
SD	2192	513	1273	3759
CV	0.20	0.03	0.12	0.29

Relationship between Maze Speed Measures and EI Performance

Raw data analysis

Table 3 presents Pearson's correlation coefficients between the EI performance and the speed measures yielded by the analysis of raw data pertaining to the maze task. The coefficient *r* was interpreted based on the guidelines for L2 research proposed by Plonsky and Oswald (2014): small–weak (≈.25), medium–moderate (≈.40), and large–strong (≈.60). In the word-level analysis, all EI measures were negatively correlated with RT and SD with a medium effect size ($-.572 < r < -.369$, $p < .01$), whereas CV was not related to any of the EI measures ($-.136 < r < -.066$, $p > .05$). The correlation between RT and CV was negligible ($r = .112$, $p = 261$).

The sentence-level analysis revealed that the students' EI performance was moderately related to both RT and SD ($-.573 < r < -.416$, $p < .01$), which is consistent with the word-level analysis. CV was negatively related to all EI performance measures with a small effect size ($-.245 < r < -.211$, $p < .05$), which suggests that CV is more strongly related to the EI performance at the sentence level than at the word level. However, the correlation between RT and CV was again negligible ($r = .189$, $p = .057$).

Table 3. *Results from correlational analyses between EI performance and speed measures on the maze task (Raw data analysis).*

	Word-Level Analysis				Sentence-Level Analysis			
	RT	SD	CV	RT-CV	RT	SD	CV	RT-CV
Repetition	-.572**	-.519**	-.125	.112	-.573**	-.525**	-.230*	.189
Both Morph.	-.495**	-.454**	-.112		-.490**	-.479**	-.245*	
Third Person	-.424**	-.369**	-.066		-.416**	-.429**	-.239*	
Plural	-.480**	-.458**	-.136		-.477**	-.447**	-.211*	

A multiple regression analysis was conducted, for parsimony, only when both RT and CV were significantly related to the EI performance. Two separate multiple regression analyses were conducted of the EI repetition accuracy and the morphological accuracy for both structures, with RT and CV from the sentence-level analysis serving as the predictors. The omnibus test revealed that both models were significant, with repetition accuracy, F (2, 99) = 25.932, $p < .001$, $R^2 = .344$, and with morphological accuracy, F (2, 99) = 17.741, $p < .001$, $R^2 = .264$.[5] RTs were significant predictors ($\beta = –.549$, $p < .001$ and $\beta = –.460$, $p < .001$ for repetition and morphological accuracy, respectively), whereas CVs were not ($\beta = –.062$, $p = .459$ and $\beta = –.158$, $p = .075$, respectively).

Cleaned data analysis

Because the authors of previous studies analyzed their RT data after cleaning, we likewise conducted a cleaned data analysis. Table 4 presents Pearson's correlation coefficients between the EI measures and the cleaned speed measures of the maze task. The word-level analysis revealed that RT and SD showed a weaker relation to the EI performance than in the raw data analysis ($–.324 < r < –.160$). None of the EI measures were related to CV ($–.099 < r < .086$, $p > .05$) and the correlation between RT and CV was almost zero ($r = –.039$, $p = .696$).

In the sentence-level analysis, the magnitudes of correlations were similar to those in the raw data analysis ($–.597 < r < –.371$). However, CV, was no longer related to any of the EI measures ($–.009 < r < –.161$). The correlation between RT and CV was negligible ($r = .041$, $p = .703$). No multiple regression analysis was conducted, because CV was not related to any of the EI measures.

Table 4. *Results from correlational analyses between EI performance and speed measures on the maze task (cleaned data analysis).*

	Word-Level Analysis				Sentence-Level Analysis			
	RT	SD	CV	RT-CV	RT	SD	CV	RT-CV
Repetition	-.324**	-.296**	-.038	-.039	-.597**	-.415**	-.009	.041
Both Morph.	-.297**	-.257**	-.013		-.488**	-.413**	-.114	
Third Person	-.255**	-.160	.086		-.396**	-.387**	-.161	
Plural	-.287**	-.303**	-.099		-.493**	-.371**	-.053	

Role of Immersion Experience

Raw data analysis

Since a prolonged stay in an English-speaking country may facilitate automatization, the previously described analyses were conducted separately for participants that reported having prior immersion experience (Experienced learners, $n = 33$) and those that did not (EFL learners, $n = 69$). Table 5 presents Pearson's correlation coefficients between the EI performance and the speed measures from the raw data analysis in the maze task separately for these two groups.

Table 5. *Results from correlational analyses among EI performance and speed measures on the maze task (Raw data analysis): Experienced learners versus EFL learners.*

	Word-Level Analysis				Sentence-Level Analysis			
	RT	SD	CV	RT-CV	RT	SD	CV	RT-CV
Experienced Learners (n = 33)								
Repetition	-.554**	-.469**	-.166	.419*	-.534**	-.526**	-.379*	.592**
Both Morph.	-.535**	-.390*	-.040		-.508**	-.503**	-.366*	
Third Person	-.450**	-.316	-.014		-.431*	-.436*	-.328	
Plural	-.548**	-.410*	-.058		-.517**	-.503**	-.358*	
EFL Learners (n = 69)								
Repetition	-.559**	-.527**	-.094	-.029	-.571**	-.536**	-.212	.077
Both Morph.	-.441**	-.461**	-.154		-.449**	-.457**	-.224	
Third Person	-.364**	-.356**	-.087		-.363**	-.410**	-.233	
Plural	-.413**	-.454**	-.180		-.427**	-.399**	-.168	

The word-level analysis revealed consistent patterns in the findings for both groups. More specifically, while RT and SD were significantly and negatively correlated with almost all EI measures with a small to medium effect size ($-.559 < r < -.316$), CV was not significantly related to any measures ($-.180 < r < -.014$). Strikingly, the correlation between RT and CV was positive and significant only in the group that had prior immersion experience ($r = .419$, $p = 015$).

The sentence-level analysis revealed a level of correlation between RT and SD and all EI measures that was similar to the level found in the word-level analysis ($-.571 < r < -.363$). Correlations between CV and the EI measures were stronger in the group with prior immersion experience ($-.379 < r < -.328$) than in the group with no such experience ($-.233 < r < -.168$). The correlation between RT and CV was significant with a large effect size ($r = .592$, $p < .001$) in the group with prior immersion experience only.

Since the sentence-level analysis produced RT and CV that were significantly correlated with EI repetition and morphological accuracy in the group with prior immersion experience, two multiple regression analyses were conducted of repetition and morphological accuracy with RT and CV as predictors. The omnibus test revealed that both models were significant, with repetition accuracy, F (2, 30) = 6.178, $p = .006$, $R^2 = .292$, and morphological accuracy, F (2, 30) = 5.391, $p = .010$, $R^2 = .264$. RTs were significant predictors ($\beta = -.477$, $p = .018$ and $\beta = -.448$, $p = .028$ for repetition and morphological accuracy, respectively); however, CVs were not ($\beta = -.097$, $p = .616$ and $\beta = -.101$, $p = .607$, respectively).

Cleaned data analysis

After data cleaning, the overall correlation coefficients were lower than those yielded by the raw data analysis (Table 6). In the word-level analysis, most relationships between RT, SD, and the EI measures were weak ($-.365 < r < -.115$), with only a few that reached statistical significance. CV was not related to any EI measures in either group, and the RT–CV correlation was no longer significant for either group ($r = -.066, -.003$, $p = .716, .979$).

Table 6. *Results from correlational analyses among EI performance and speed measures on the maze task (cleaned data analysis): Experienced learners versus EFL learners.*

	Word-Level Analysis				Sentence-Level Analysis			
	RT	SD	CV	RT-CV	RT	SD	CV	RT-CV
Experienced Learners (n = 33)								
Repetition	-.278	-.234	-.033	-.066	-.429*	-.174	.155	.195
Both Morph.	-.318	-.249	-.017		-.388*	-.152	.140	
Third Person	-.224	-.164	.006		-.314	-.139	.093	
Plural	-.365*	-.295	-.036		-.417*	-.148	.171	
EFL Learners (n = 69)								
Repetition	-.271*	-.302*	-.116	-.003	-.622**	-.477**	-.056	-.035
Both Morph.	-.227	-.232	-.061		-.465**	-.498**	-.231	
Third Person	-.200	-.115	.101		-.351**	-.465**	-.288*	
Plural	-.201	-.284*	-.189		-.460**	-.416**	-.134	

In the sentence-level analysis, RTs were correlated with the EI measures more strongly than in the word-level analysis. In addition, while RT was negatively correlated with the EI measures with small to medium effect sizes across the groups ($-.622 < r < -.314$), SD was unexpectedly not related to the EI for immersion in the group with prior immersion experience ($-.174 < r < -.139$). Critically, no CVs were significantly related to any of the EI measures ($p > .05$), with the exception of a weak association between CV and accuracy in third person *s* ($r = -.288$, $p = .026$). Once again, the RT–CV correlation was no longer statistically significant for either group ($r = .195, -.035$, $p = .312, .790$).

Discussions

Quantitative versus Qualitative Development of Automatization

This study investigated to what extent RT and CV contributed to L2 oral proficiency. The RT and CV were measured by the maze task, and the EI task assessed oral proficiency in terms of repetition accuracy and the use of third person *s* and plural *s*. First, RT was negatively correlated with all EI performance measures with medium effect sizes, which suggests that faster RT leads to better EI performance. Secondly, no correlation between RT and CV was found, which indicates that automatization was not achieved in the entire study sample. Thirdly, CV was not correlated with any EI performance measures. The multiple regression analyses confirmed that RT, not CV, was the only significant predictor of L2 oral proficiency.

These findings suggest that a group of EFL learners that received six years of English instructions showed virtually no evidence of genuine automatization. In addition, RT (processing speed) was a stronger predictor of their EI performance than CV (processing stability). In lexical access, both RT and CV were found to be significantly correlated with oral fluency in the interview task (e.g., Segalowitz & Freed, 2004). At the sentence level, however, the utility of CV seems limited. This assertion supports the findings reported by Hulstijn et al. (2009) and

contradicts the more recent findings published by Lim and Godfroid (2015), which raises the question of the usefulness of CV. We concur with Hulstijn et al.'s (2009) claim that "we wonder whether a mathematical distinction so subtle should be taken as forming the empirical litmus test for a conceptual distinction so important" (p. 579). Simply being faster in syntactic processing (i.e., mere speedup) may be important enough to improve L2 oral performance in the EFL setting.

Another interpretation of the limited role of genuine automatization of sentence processing in oral proficiency may pertain to skill specificity. In the current maze task, the participants *read* sentences, and their RTs were compared with *speaking* proficiency. It has been found that highly automatized skills are unlikely to transfer to other domains of skills (DeKeyser, 1997; DeKeyser & Sokalski, 1996). In light of this skill-specific view of transfer, it is reasonable to assume that automaticity in reading may not transfer well to speaking performance. Therefore, we cannot draw the definitive conclusion that CV is not a useful index to measure L2 learners' automaticity in sentence processing. Possibly, CV of sentence processing in reading can be a significant predictor of reading comprehension skills.

This study focused on a group of L2 learners whose L1 (Japanese) is typologically different from English. After comparing our findings with those from two previous studies (Hulstijn et al., 2009; Lim & Godfroid, 2015), the typological difference between L1 and L2 does not seem to explain the discrepancy between the results (L1 Dutch versus L1 Korean). Since the Japanese and Korean languages have very similar L1 syntactic structures, which are very different from those found in English, the development of automaticity could have been better observed among Japanese learners due to typological difference. However, the Japanese learners in the present study showed very little evidence of automatization.

Role of Immersion Experience in Automatization

An intriguing pattern of findings emerged when learners were divided into two groups based on experience in immersion settings. While no RT–CV correlation was found among those that had never studied abroad in an English-speaking country, a positive RT–CV correlation was found among L2 learners with immersion experience (r = .592, based on the sentence-level analysis). L2 learners who had studied English in an EFL context and had been immersed for at least one month showed some indication of genuine automatization, which suggests that restructuring occurred (e.g., elimination and reorganization of unnecessary processing).

The context or role of immersion experience in English-speaking countries seems to be a crucial factor for the attainment of automatization. In skill acquisition theories, a large amount of practice is assumed necessary for attaining automaticity. In particular, a stay abroad provides an ideal environment for automatization (DeKeyser, 2007). Therefore, tests that assess L2 automaticity may be better suited for learners who have some immersion experience, as they are more likely to attain automaticity through ample exposure. In other words, L2 learners who studied abroad might have been at a later, automatization stage of skill acquisition, in which they started automatizing their declarative knowledge. In contrast, L2 learners, who had never experienced immersion environments, were probably still at an earlier, declarative learning stage of skill acquisition (e.g., learning about grammatical rules). Since mere speedup precedes automatization, RTs, and not CVs, could be more powerful predictors of L2 speaking proficiency, particularly for L2 learners without immersion experience.

This experience factor can explain the discrepancy between the findings reported by Hulstijn et al. (2009) and Lim and Godfroid (2015), respectively. The L2 learners in the study

conducted by Hulstijn et al. (2009) were EFL learners with limited immersion experience and at an earlier stage of development, who showed very little evidence of automatization. In contrast, the L2 learners that took part in Lim and Godfrod's (2015) research were enrolled at a university in the United States and had numerous and frequent opportunities to engage in extensive practice, which facilitated automatization, as their study indicated. Using the same set of EI and maze tasks, the current study has demonstrated that immersion experience plays a crucial role in the automatization process, as measured by the maze task.

Although L2 leaners with immersion experience showed some indication of automatization, even in this group, CV itself was not proven to be a better predictor of oral proficiency than RT. Even if these learners managed to eliminate extraneous processing, this caused almost no change in their overt, e.g., oral performance in L2. In sum, the present findings lead us to question the usefulness of CV as an index for automatization.

Advantages and Disadvantages of Using Maze Task as a Sentence Processing Measure

In previous studies in this field, researchers used the sentence construction task to assess sentence processing speed. In contrast, the current study employed the maze task, which can be regarded as its extension. We believe that the maze task can be more advantageous than the sentence construction task, because it can examine the processing of an entire sentence, thus achieving more ecological validity in capturing sentence construction processes. As a case in point, the maze task has been found to provide a pattern of reading processes that is similar to those found by means of more naturalistic data collection methods, such as eye-tracking technique (Witzel et al., 2012). In the sentence construction task, participants only need to read the first phrase and only once choose the correct word that continues the phrase. When participants perform the maze task, they have to actually build a sentence from beginning to the end. This means that participants have to constantly retain and integrate previous information and build the sentence incrementally. Incremental and predictive sentence processing is regarded as a hallmark of rapid sentence processing (Altmann & Kamide, 2007; Kamide, Scheepers, & Altmann, 2003).

Despite its advantages, the maze task entails some drawbacks. Firstly, the maze task can never assess natural reading behavior, as participants have to read sentences in fragments. Secondly, the maze task is more likely to induce noise in RTs every time a participant presses a button, which is partly due to inter-individual variation in physical response. In other words, RTs from a maze task that requires participants to press a button more than 5 times could contain greater noise. Sentence construction, however, requires only one response for each test item, which may reduce the variance in errors due to individuals' physical response. Thirdly, in a maze task, participants may sometimes have to use semantic information (e.g., background information) while reading an entire sentence. This comprehension process cannot be automatized; only lexical retrieval and syntactic processing can be automatized through practice (see further discussion in Lim & Godfroid, 2015). The requirement of reading an entire sentence might have made it more difficult for participants of a maze task to tap into automatic processing skills.

Some Considerations on Data Analysis of Maze Task

Since the maze task is purported to capture the reading speed of an entire sentence, it seems that the analysis of RT data should be consistent with that view. We analyzed RT both at the word level and at the sentence level, and the results consistently showed that the sentence-level analysis provided stronger associations between RT, CV, and the EI measures than the word-level analysis did. The sentence-level analysis was proven to be superior, probably because speed measures can reflect the processing speed of a sentence as a whole. Note that the word-level analysis may have been advantageous because a larger number of RT data points were available for analyzing RT variability (CV) more stably. The current evidence suggests that a sentence-level analysis is recommended, which reflects the view that the maze task is suitable for assessing sentence processing skills.

Another issue related to the analysis of RT tasks stems from the difference between the raw data and cleaned data analyses. In the present study, the raw data analysis consistently showed stronger associations between RT, CV, and the EI scores. However, Lim and Godfroid reported that data cleaning had almost no impact on the results of the sentence construction task, as the accuracy score was near the 100% ceiling (98%). In contrast, the accuracy scores were lower (87% and 77%, at the word- and the sentence-level, respectively) for the current maze task. More particularly, about 14% of participants and 22% of items were excluded from the sentence-level analysis. These data cleaning procedures seemed to have contributed to the different pattern of findings observed in the raw-data analysis.

This study employed the data cleaning procedure to isolate RT and CV from the differences in the accuracy rate. In theory, automatization can be distinguished from accuracy development (Hulstijn et al., 2009); however, accuracy and RT (CV) develop simultaneously, and data cleaning may artificially and incorrectly affect both. Concurring with Lim and Godfroid (2015), the present findings suggest that a raw data analysis may be better than a cleaned data analysis. Since, both the difficulty of the task and the proficiency level of L2 learners influence the type of responses that are gathered, researchers need to examine how these two types of analysis influence the validity of RT and CV.

Suggestions for Future Research

The present study is not without limitations and opens up new avenues for further research. This study employed a cross-sectional design. As Lim and Godfroid (2015) pointed out, two assumptions need to be met for between-subject experiments. First, all L2 learners are assumed to experience common automatization processes, which are supported by L2 skill acquisition theories (DeKeyser, 2015; McLaughlin, 1987; N. S. Segalowitz & S. J. Segalowitz, 1993). Secondly, the development of L2 proficiency needs to reflect a higher level of automaticity. Since we used the EI task as a proficiency measure, it is safe to assume that participants' EI scores reflect automaticity, because the task imposed time pressure on processing and imitation in order to draw on the participants' more automatic linguistic knowledge (Erlam, 2006; Spada et al., 2015; Suzuki & DeKeyser, 2015). Nevertheless, there is a critical need for a longitudinal design to further investigate the automatization process through RT and CV that may potentially distinguish speedup and automatization. CV might not have been sensitive enough in the between-subject design, as RT is standardized for CV (i.e., CV indicates RT variability independently of the absolute differences in RTs). A study adopting a within-subject approach is needed.

Since syntactic or sentence processing consists of several cognitive processes, such as lexical retrieval, the maze task did not provide a pure measure of syntactic processing. Although most words in stimulus sentences were familiar to participants, lexical retrieval speed might have been a confounding factor. A task that specifically aims to assess lexical access speed (e.g., a semantic classification task; see Lim & Godfroid, 2015) should therefore be employed in conjunction with a sentence-level task.

The current study used the EI task as a measure of oral proficiency. Although the EI task provides a good estimate of speaking fluency, it does not provide temporal fluency measures of speech, such as syllables per minute or length of pauses. It may be worthwhile to employ a more unconstrained, free speaking task and analyze temporal measures of utterances (e.g., De Jong & Perfetti, 2011).

In addition, the EI task might not have been an ideal task to assess L2 proficiency, as the task was very difficult for the current sample of learners *(M* = 1.60 out of 4). This led to a small variation in the score of the EI task (*SD* = .62), which could have limited the utility of CV. Furthermore, the learners' immersion experience varied greatly between individuals, as indicated by the SD, which was larger than the mean length of stay in months (*M* = 15 months, *SD* = 24 months). This suggests that the learning experience that contributed to automatization was different within the study-abroad group. The present findings should be attested for different samples of L2 learners, by isolating relevant factors (e.g., length of stay).

Conclusions

The present study set out to investigate to what extent quantitative speedup (RT) and qualitative change of syntactic processing or genuine automatization (CV) were related to L2 oral performance The results show that Japanese EFL learners exhibited very limited automatization. Immersion experience seemed to influence the automatization process, which suggests that restructuring in L2 sentence processing may actually occur after the immersion experience. However, the sheer speedup indexed by RT was a better predictor of L2 oral proficiency than CV, irrespective of immersion experience. Our tentative conclusions restrict the utility of CV as an index of automatization in L2 sentence processing. We hope that the present study will stimulate further interest in the assessment of automatization in L2 processing

Appendix A. Information about the stimulus sentences for lists 1 and 2.

	Num. of Words	Num. of Syllables	Speech Length in Second	Words Per Minute	Syllable Per Minute
List 1	10.38 (8-13)	13.79 (9-18)	3.58 (2.47-3.58)	176.04 (131.71-235.29)	231.93 (165.93-302.84)
List 2	10.52 (7-13)	13.60 (9-18)	3.54 (2.34-4.53)	180.24 (117.65-230.77)	230.69 (181.82-269.13)

Note. The numbers in the bracket indicate the range (minimum-maximum).

Appendix B. Data cleaning results for the maze task.

Criteria	Percentage	Word-level Analysis	Sentence-level analysis
(1) Participant's accuracy < 62.5%	Exclusion	1.96% (2/102)	13.73% (14/102)
(2) Item accuracy < 75%	Exclusion	12.72% (115/904)	21.88% (21/96)
(3) Outlying RTs < 300 ms, > 3SD + Mean	Missing values	1.28% (595/46616)	6.03% (208/3449)
(4) Multiple Imputation	Imputed cases	10.73% (4326/40311)	515/3405 (15.12%)

Appendix S1: Stimulus sentences in Lists 1 and 2

List 1

1. The population of the world increases every year.
2. My sister usually gets to school at eight in the morning.
3. The dog over there runs faster than any other dogs.
4. My brother always eats breakfast in the morning.
5. The English teacher always asks the students to read difficult words.
6. My brother works hard to make a lot of money.
7. Mary always gives food to her rabbit when the rabbit is hungry.
8. When it is raining, John goes to the library near his house.
9. John often rides his bicycle to work when it is sunny.
10. If the girl studies hard, she can watch TV.
11. Your health becomes worse if you eat too much.
12. When your child catches a cold, you should go to the hospital.
13. Where are you going to put all the old pictures?
14. How did the lady manage to learn four different languages?
15. When did the young woman start wearing glasses?
16. How much did you pay for all the textbooks this year?
17. How long did it take to cook all the dishes?
18. How many times did you visit the museum in three years?
19. Where did the young man find ten birds today?
20. When did you learn to play so many songs on the piano?
21. Why do we have twenty chairs in the kitchen?
22. What did your father buy for his three kids on Christmas?
23. Why do you have two soccer balls at home?
24. What do your five kids usually eat for breakfast?
25. The boy who is kissing the girl goes to a famous school.
26. The doctor who is kind to the family lives in London.
27. The musician who will have a concert tomorrow plays the piano.
28. The lady knows the shop which is popular in Tokyo.
29. The mother knows a lady who can speak French and Spanish.
30. The policeman talks to boys who do not go to school.
31. The boy whom you like best plays basketball very well.
32. The tennis racket which Mary found costs fifty dollars.
33. The girl whom I know opens a new shop.
34. John's sister goes to a university which many people know.
35. The farmer sells vegetables which many old people can enjoy.
36. The boy loves the bike which he bought three years ago.
37. The boy wondered if he should take three classes at school.

38. The girl wondered if she should have two hamburgers at McDonald's.
39. The teacher knew whether some students could do well on the test.
40. A lot of foreigners asked whether the train was going south.
41. Many students wondered if the school would be closed tomorrow.
42. My teacher knew whether many students would join the club.
43. Many students asked if the exam would be difficult.
44. The woman wondered if most cars would be more expensive.
45. The father wanted to know whether his two kids would go to college.
46. The boy asked if he could see many trains at the station.
47. The mother wondered if her son is dating two girls.
48. The mother needed to know if the father would buy three cars.

List 2

1. The population of Japan decreases every year.
2. The famous scientist works every day to improve our lives.
3. The famous dancer performs next Sunday at a large theater.
4. The red flower on the table looks so beautiful.
5. My teacher often asks a lot of difficult questions in class.
6. The female student often takes a train to school.
7. My mother always tells me to study because I do not study.
8. My mother sings a song loudly when she is happy.
9. The cat often comes to my house when it is raining.
10. If the boy studies hard, he can play baseball.
11. When my father cooks dinner, I have to help him.
12. If you laugh a lot, your health gets better
13. Where can I buy five notebooks for my classes next semester?
14. How can we solve many difficult questions without help?
15. Why do you want to watch two movies in one day?
16. How much did it cost for twelve eggs at the store?
17. How long did it take to write five long stories?
18. How many times do you wash your hands?
19. Where did your father find three cats last month?
20. When did your teacher buy so many books for you?
21. What do most of the high school students do after school?
22. What did you do with some classmates this weekend?
23. Why does your father keep ten dogs in his house?
24. What did Mary buy at the department store two years ago?
25. The student who is good at speaking French studies hard
26. The man who can play the guitar well likes rock music.
27. The boy who can run fast catches a cold every winter.

28. My mother knows the restaurant which is famous for excellent food.
29. The mother likes the artist who can paint beautiful pictures.
30. The teacher always supports students who are interested in English.
31. The pianist whom you like best joins the concert at the festival.
32. The interesting book which many students like costs one thousand yen.
33. The actor whom I really like lives in a big house.
34. My sister reads English books which she can borrow from the library.
35. Daniel tries to sell a car which he used for two years.
36. The boy likes the pictures which he took at the zoo.
37. Some girls wondered if they should go to school every day.
38. You asked if I would like to eat two cakes at the restaurant.
39. The teacher knew whether most of the students could pass the test.
40. Many students wondered whether they could do well at the contest.
41. The boy wondered if his four dogs could live longer.
42. The man knew whether his three kids could come to the party.
43. The young man wondered if he should get two cats this year.
44. The actress wondered if some of her movies would be popular.
45. My mother asked if I have many good friends at school.
46. The mother asked if her daughter would like to buy more books.
47. Most of my friends knew if I would take the English course.
48. The mother knew if the father bought many books.

References

Akamatsu, N. (2008). The effects of training on automatization of word recognition in English as a foreign language. *Applied Psycholinguistics, 29*(02), 175-193. doi:doi:10.1017/S0142716408080089

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. (2015). *Cognitive psychology and its implications* (8th ed.). New York, NY: Worth Publishers.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior, 14*(6), 575-589.

Chapelle, C. A., & Heift, T. (2009). Individual Learner Differences in CALL: The Field Independence/Dependence (FID) Construct. *CALICO journal, 26*(2), 246-266.

Cobb, T. (2002). Web Vocabprofile. Retrieved from http://www.lextutor.ca/vp/

Cowan, N., Day, L., Saults, J. S., Keller, T. A., Johnson, T., & Flores, L. (1992). The role of verbal output time in the effects of word length on immediate memory. *Journal of memory and language, 31*(1), 1-17.

De Jong, N., & Perfetti, C. A. (2011). Fluency Training in the ESL Classroom: An Experimental Study of Fluency Development and Proceduralization. *Language Learning, 61*(2), 533-568.

De Jong, N., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34*(5), 893-916.

DeKeyser, R. M. (1997). Beyond Explicit Rule Learning. *Studies in second language acquisition, 19*(2), 195-221.

DeKeyser, R. M. (2007). Study abroad as foreign language practice. In R. M. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 208-226). New York, NY: Cambridge Universtiy Press.

DeKeyser, R. M. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 94-112). New York, NY: Routledge.

DeKeyser, R. M., & Sokalski, K. J. (1996). The Differential Role of Comprehension and Production Practice. *Language Learning, 46*(4), 613-642. doi:10.1111/j.1467-1770.1996.tb01354.x

Ellis, R. (2009). Investigating Learning Difficulty in Terms of Implicit and Explicit 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. 139-142). Tonawanda, NY: Multilingual Matters.

Enkin, E. (2012). The maze task: Training methods for second language learning. *Arizona Working Papers in SLA & Teaching, 19*, 56-81.

Enkin, E., & Forster, K. I. (2014). Examining the Training Effect of Using a Psycholinguistic Experimental Technique for Second Language Learning. *JLLT, 5*(2), 161-180.

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

Europe, C. o. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment.(CEFR). New York: Cambridge University Press.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*(1), 116-124.

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods, 41*(1), 163-171.

Harrington, M. (2006). The lexical decision task as a measure of L2 lexical proficiency. *EUROSLA Yearbook, 6*(1), 147-168.

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.

Jensen, E. D., & Vinther, T. (2003). Exact Repetition as Input Enhancement in Second Language Acquisition. *Language Learning, 53*(3), 373-428. doi:10.1111/1467-9922.00230

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.

Koda, K. (2007). Reading and Language Learning: Crosslinguistic Constraints on Second Language Reading Development. *Language Learning, 57*(s1), 1-44.

Kormos, J. (2006). *Speech production and second language acquisition*. New York: Routledge.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.

Levelt, W. J. M. (1999). Producing spoken language: A blueprint of the speaker. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83-122). Oxford, UK: Oxford University Press.

Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics, 36*(5), 1247-1282. doi:doi:10.1017/S0142716414000137

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah: NJ: Routledge.

McBride, K. (2011). The effect of rate of speech and distributed practice on the development of listening comprehension. *Computer Assisted Language Learning, 24*(2), 131-154.

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

Ortega, L., Iwashita, N., Norris, J., & Rabie, S. (2002). *An investigation of elicited imitation in crosslinguistic SLA research.* Paper presented at the Conference handout from paper presented at the meeting of the Second Language Research Forum, Toronto, Canada.

Plonsky, L., & Oswald, F. L. (2014). How Big Is "Big"? Interpreting Effect Sizes in L2 Research. *Language Learning, 64*(4), 878-912. doi:10.1111/lang.12079

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.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association, 72*(359), 538-543.

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

Segalowitz, N. S. (2010). *Cognitive bases of second language fluency*. NY: Taylor & Francis.

Segalowitz, N. S., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition. *Studies in second language acquisition, 26*(2), 173-199.

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

Segalowitz, N. S., Watson, V., & Segalowitz, S. J. (1995). Vocabulary skill: single-case assessment of automaticity of word recognition in a timed lexical decision task. *Second Language Research, 11*(2), 121-136. doi:10.1177/026765839501100204

Segalowitz, S. J., Segalowitz, N. S., & Wood, A. G. (1998). Assessing the development of automaticity in second language word recognition. *Applied Psycholinguistics, 19*(1), 53-67.

Spada, N., Shiu, J. L.-J., & Tomita, Y. (2015). Validating an Elicited Imitation Task as a Measure of Implicit Knowledge: Comparisons With Other Validation Studies. *Language Learning, 65*(3), 723-751. doi:10.1111/lang.12129

Suzuki, Y., & DeKeyser, R. M. (2015). Comparing Elicited Imitation and Word Monitoring as Measures of Implicit Knowledge. *Language Learning, 65*(4), 860-895. doi:10.1111/lang.12138

Tracy-Ventura, N., McManus, K., Norris, J., & Ortega, L. (2014). 'Repeat as much as you can': Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, A. AEdmonds, & H. Hilton (Eds.), *Measuring L2 Proficiency: Perspectives from SLA*. *Bristol: Multilingual Matters* (pp. 143-166).

Witzel, N., Witzel, J., & Forster, K. I. (2012). Comparisons of Online Reading Paradigms: Eye Tracking, Moving-Window, and Maze. *Journal of psycholinguistic research, 41*(2), 105-128. doi:10.1007/s10936-011-9179-x

Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals, 46*(4), 680-704.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2015). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, Online First*.

Footnotes

[1] In the present paper, we use sentence processing and syntactic processing interchangeably, while making a distinction from lexical processing.

[2] Since no independent proficiency measure was administered to the Dutch students, it is hard to assess whether their proficiency actually developed during those two years. It may be the case that the Dutch high school learners did not progress in gaining declarative knowledge or failed to automatize their knowledge during those two years.

[3] This value is based on the previous study that employed the maze task (Enkin & Forster, 2014).

[4] The study conducted by Lim and Godfroid (2015) compared mean RTs of the intermediate and advanced L2 proficiency groups. The two groups were created based on L2 proficiency, as indicated by vocabulary and grammar test scores. Our correlational approach can examine a linear relationship between RT and L2 proficiency without dividing the L2 groups somewhat arbitrarily based on the test scores.

[5] Assumptions of the absence of multicollinearity were met, with VIF < 10 and tolerance > .02 (Chapelle & Heift, 2009). The remaining analyses met these assumptions as well.