This is an accepted manuscript published in *Language Learning*. Please cite as:

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

Comparing Elicited Imitation and Word Monitoring as Measures of Implicit Knowledge

Yuichi Suzuki and Robert DeKeyser

University of Maryland

The present study challenges the validity of elicited imitation (EI) as a measure for implicit knowledge, investigating to what extent the online error detection and the subsequent sentence repetition draw on implicit knowledge. To assess the online detection during listening, a word-monitoring component was built into the EI task. Advanced-level Japanese L2 speakers with Chinese as their native language performed the EI task with the built-in word-monitoring component, a metalinguistic knowledge test, and a probabilistic serial reaction time (SRT) task which served as a measure of aptitude for implicit learning. Results showed that the EI scores were correlated positively with metalinguistic knowledge, but they were not related to the SRT scores. The word-monitoring performance, in contrast, was not related to metalinguistic knowledge but correlated positively with the SRT scores only among L2 speakers with longer lengths of residence. These results suggest that online error detection can index implicit knowledge, whereas EI may measure automatized explicit knowledge.

Keywords elicited imitation; word monitoring task; explicit/implicit knowledge; serial reaction time task; metalinguistic knowledge

Introduction

"There are good theoretical and educational reasons to place matters of implicit and explicit learning high on the agenda for SLA research" (Hulstijn, 2005, p. 129). Many second language (L2) acquisition researchers have been interested in the products of these two distinct learning mechanisms and how they eventually result in explicit and/or implicit knowledge, that is, the so-called "interface issue" (DeKeyser, 2003; N. C. Ellis, 2005; R. Ellis, 2005; Krashen, 1981; Paradis, 2009). The methodological problem of measuring implicit knowledge (i.e., use of linguistic knowledge without awareness) is critical for tackling these issues: Valid and fine-grained measurements for explicit and implicit knowledge are needed. The aim of the current study is to validate two techniques that have been suggested in the literature as measures of implicit knowledge, an elicited imitation (EI) task (Bowles, 2011; R. Ellis et al., 2009; Erlam, 2006; Zhang, 2014) and a word-monitoring task (Granena, 2013a; Jiang, 2011; Jiang, Hu, Lukyanchenko, & Cao, 2010).

Background

Elicited Imitation

EI tasks have been extensively used in first language (L1) acquisition research to assess L1 development (e.g., Fraser, Bellugi, & Brown, 1963) and have also received attention from L2 researchers (Bley-Vroman & Chaudron, 1994; R. Ellis, 2005; Erlam, 2006; Jessop, Suzuki, & Tomita, 2007; Vinther, 2002). A number of researchers agree that EI taps linguistic knowledge (Bley-Vroman & Chaudron, 1994); controversy exists, however, about exactly what psycholinguistic construct it measures. This issue is of particular interest to researchers because the question of whether L2 competence consists of implicit knowledge, explicit knowledge, or a

combination of both (Jessop et al., 2007) is central to many issues in the field, and therefore having valid operational measures of both is crucial.

Several L2 acquisition researchers have attempted to validate EI as a measure of implicit knowledge (Bowles, 2011; R. Ellis, 2005; Erlam, 2006; Zhang, 2014). In these studies, the EI task is used to assess the grammatical knowledge of specific L2 structures. In order to be valid as an implicit knowledge measure, EI must (a) be reconstructive and (b) draw on implicit knowledge exclusively. Researchers have made efforts to make EI reconstructive, that is, to design the test so that it requires participants to reconstruct the stimulus sentence with their own grammar. This is necessary to support the claim that EI measures linguistic knowledge or L2 competence rather than short-term memory capacity. The reconstructive process is a necessary, but not a sufficient condition for measuring implicit knowledge, however. In the following sections, we first review how EI procedures can be tailored to be reconstructive. Next, criteria for implicit knowledge measures are critically evaluated.

Elicited Imitation as Reconstructive Task

EI was once criticized for being a measure of short-term memory capacity or rote memory ability without comprehension (Fraser, Bellugi, & Brown, 1963). There is research evidence, however, to suggest that if the task is carefully constructed, EI can assess linguistic knowledge independently of short-term memory. In order to make EI reconstructive (i.e., so that it does not merely involve retrieval of information from memory), the EI should be designed (see Mackey & Gass, 2005, for details) such that it can (a) direct participants' attention to meaning; (b) include a delay between the stimuli presentation and repetition to prevent rote repetition; (c) include sentences long enough to exceed short-term memory capacity (Baddeley, Thomson, & Buchanan, 1975); and (d) embed the target structure in the middle of the sentence (Bley-Vroman &

Chaudron, 1994). In addition, Erlam (2006) added another element to the technique, namely, including stimulus sentences with grammatical errors to see whether participants would repeat or correct the error. As this new element, along with the instructions it necessitates (showing modeled responses that do not repeat the errors), may draw participants' attention to form, Erlam added a second element: comprehension (belief-statement) questions in order to reduce focus on form.

In addition to these design features, a number of criteria can also be applied to the outcome measures to assess whether test-takers actually reconstructed the sentence. Among them, the current study evaluates two criteria for reconstructive processing: (a) correction of ungrammatical sentences, which can be demonstrated by a high positive correlation between the scores in grammatical and ungrammatical sentences (Erlam, 2006) and (b) the lack of correlation between the success of imitation and short-term or working memory capacity (Okura & Lonsdale, 2012). If the participants simply use short-term memory storage for EI performance, they are less likely to correct the ungrammatical part of the sentence but they should succeed in repeating the grammatical sentences as they are, yielding no correlation between the scores for the grammatical and ungrammatical items. If, in contrast, they draw on linguistic knowledge to reconstruct the ungrammatical sentence, the scores for grammatical and ungrammatical items should correlate with each other. If the EI score is not a function of the individual's memory capacity, this adds further confidence that participants are performing the task using their linguistic knowledge rather than their memory of the stimuli. Reconstructive processing of EI is a necessary condition to measure implicit knowledge, but it is not sufficient.

Attention, Awareness and Noticing During Elicited Imitation

The issue of awareness is central to the discussion of implicit and explicit learning (see DeKeyser, 2003; Williams, 2009, for review). Schmidt (2001) distinguished two types of awareness, that is, awareness at the level of noticing and awareness at the level of understanding (i.e., metalinguistic awareness). Tomlin and Villa (1994) further proposed a more fine-grained analysis of the attention mechanism, as attention is seen as controlling access to awareness (Posner, 1994). They claimed that the attention system consists of three separate, but related processes: alertness, orientation and detection. Alertness refers to the readiness to process information, while orientation refers to the allocation of attention on an incoming stimulus. Detection refers to "cognitive registration of sensory stimuli" and "the process that selects, or engages, a particular and specific bit of information" (Tomlin & Villa, 1994, p. 192). None of these three processes requires awareness. Put another way, "awareness requires attention, but attention does not require awareness" (p. 194). Tomlin and Villa illustrated the detailed structure of the attentional system prior to awareness, and the important distinction that should be made between detection without awareness (registration) and detection within focal attention accompanied by awareness, which would correspond to conscious perception or noticing (Schmidt, 2001). In what follows, we use the terms "registration" and "detection" interchangeably to refer to cognitive processes without awareness in the restricted sense, and the terms "noticing" and "conscious perception" for cognitive processes with awareness. These constructs will be utilized in order to analyze the cognitive processes during the EI task.

Elicited Imitation as Implicit Knowledge Measure

Recent research has attempted to validate EI as an implicit knowledge measure (Bowles, 2011; R. Ellis, 2005; Erlam, 2006; Zhang, 2014). R. Ellis (2005) hypothesized seven criteria for assessment of explicit and implicit knowledge. Four of those are directly relevant to the current

study:[1] (a) degree of awareness, (b) time available, (c) focus of attention, and (d) metalinguistic knowledge. The crucial factor that differentiated between explicit and implicit knowledge measures in the four studies was time available to perform the tests (Bowles, 2011; R. Ellis, 2005; Erlam, 2006; Zhang, 2014). The factor-analytic approach used by R. Ellis and Bowles showed that the EI loaded on a factor with other time-pressured tasks (i.e., a timed grammaticality judgment task or GJT, and an oral narrative task), while the tasks that loaded on the other factors were untimed (i.e., an untimed GJT and a metalinguistic knowledge test). These results showed that the time-pressured tests were highly correlated with the EI; this was interpreted by these researchers as evidence that EI measures implicit knowledge.

Following most of the L2 acquisition literature on implicit learning, however, the present study primarily makes a distinction between explicit and implicit knowledge based on the criterion of awareness, namely, the extent to which L2 users are aware of their linguistic knowledge. The awareness criterion has been utilized in previous studies on implicit learning and in the design of measures for implicit knowledge (e.g., R. Ellis, 2005; Hama & Leow, 2010; Williams, 2005). We argue that it is superior to the time-pressure criterion because even when the task is performed under time constraints, L2 learners may still be capable of using linguistic knowledge with awareness (DeKeyser, 2003, 2009). In other words, automatized explicit knowledge can be deployed quickly under time pressure, for example, to perform an EI task. Access to explicit knowledge involves use of linguistic knowledge *with* awareness even if the execution is rapid or automatized to a high degree, which should be distinguished from the use of linguistic knowledge *without* awareness (i.e., implicit knowledge). We do acknowledge, however, that fully automatized knowledge may be used with little or no awareness and would therefore be nearly impossible to tease apart from implicit knowledge through behavioral measures. It is thus

an empirical question as to whether a behavioral measure can distinguish implicit knowledge from automatized explicit knowledge.

The issue of awareness has been tackled in some studies (e.g., R. Ellis, 2005), but a retrospective report on the EI can possibly provide a more direct way of measuring awareness. Chrabaszcz and Jiang (2014) provide some interesting data regarding the degree of awareness during the EI performance. They administered the EI task to advanced L2 English speakers to examine the acquisition of English articles. With respect to awareness, Chrabaszcz and Jiang examined whether participants noticed the grammatical errors and corrected them consciously (i.e., awareness of errors and corrections) by administering a retrospective questionnaire after the EI performance. Some of the participants became aware of the existence of grammatical errors, even though they were not told that they would hear ungrammatical sentences.

Based on the detailed constructs of awareness delineated in the previous section, we analyze the cognitive processes during the EI performance. In the processing component of EI, syntactic parsing of an auditorily presented sentence is assumed to take place primarily with attention to meaning, in order to answer comprehension questions immediately followed by the auditory stimulus. An auditory stimulus is fleeting, 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. The current study thus operationalizes implicit knowledge (i.e., language use without awareness) as registration or detection of grammatical errors in this task. The registration of errors in real time should occur without awareness, drawing on implicit knowledge.

In the production stage, in contrast, L2 speakers might be able to monitor their utterance or use automatized explicit knowledge to imitate the sentence even under time pressure because

they have enough time (e.g., several seconds) to access explicit knowledge (DeKeyser, 2003, 2009). This is partially supported by Chrabaszcz and Jiang's (2014) findings that some of the participants were aware of correcting grammatical errors during the EI task. This suggests that the registration (without awareness) of errors rises to the level of noticing or conscious perception when the input enters working memory, where maintenance rehearsal is carried out. In order to register the errors without awareness at the exact time of occurrence, implicit knowledge is necessary, however, regardless of the fact that registration can lead to further awareness.

As shown in the analyses based on the awareness criterion, the two processing components in the EI task may entail different cognitive processes. While (advanced) L2 speakers may still be able to use explicit knowledge rapidly in the production stage, there is much less room for using their automatized explicit knowledge in the listening stage. It is thus important to delve into the listening component of the EI task and examine whether L2 speakers deploy linguistic knowledge without awareness (i.e., implicit knowledge) to detect grammatical errors at the exact time of occurrence while their attention is focused on meaning. Thus far, no research has attempted to directly examine the first stage of EI, the processing of the stimulus. One technique that can test this fast deployment of implicit knowledge during the listening stage is a word-monitoring task, which we discuss in the next section.

Word Monitoring

Psycholinguistic techniques utilizing reaction time (RT) have revealed how L2 sentence processing is carried out in real time (see Jiang, 2011, for review). These methods may be useful to examine whether L2 learners can register the error without awareness while reading/listening for comprehension. Representative methods include self-paced reading tasks (Coughlin &

Tremblay, 2013; Jiang, Novokshanova, Masuda, & Wang, 2011; Roberts & Liszka, 2013) and word monitoring tasks (Granena, 2013a; Jiang et al., 2010). The present study uses a word monitoring task targeting auditory sentence processing to measure language users' ability to register grammatical errors during the processing component of the EI.

The word monitoring paradigm has been utilized previously to examine whether L2 speakers are sensitive to morphosyntactic violations (Granena, 2013a; Jiang et al., 2010). This computerized task typically includes a target word, to which participants need to respond by pressing a button as soon as they hear it. The rationale of the task is that participants slow down to respond to a target word that appears after a grammatical error, which reflects sensitivity to errors. For instance, the response time to the monitored word (i.e., *by*) in a sentence like "The book is being closely *pick by the large group of curious students" (where an asterisk designates an ungrammatical element) will be delayed when the monitoring word appears after the ungrammatical part of the sentence, compared to the grammatically correct element (i.e., *picked*). Based on this rationale, the RT difference between the grammatical and ungrammatical items, that is, the Grammatical Sensitivity Index (GSI), can indicate the extent to which participants are able to detect the error during the test. The magnitude of the index may possibly indicate how developed one's implicit knowledge is.

The word monitoring task, then, seems to hold promise for examining the registration of errors without awareness because it requires participants to engage in the dual task of monitoring the word and comprehending the sentence, which directs their attention away from form (that is, with little orientation, in the sense of the attentional framework, to ungrammatical form). The aural modality of the task is more advantageous than the written modality for examining sensitivity to errors in real time because it provides the least opportunity for reflection. As long

as the word to be monitored can appear immediately after a specific grammatical structure, and several words follow the target word, a variety of grammatical structures can be tested. The current study applied this methodology to measure online sensitivity to five Japanese particles.

Individual Differences and Implicit Knowledge

Acquisition of implicit knowledge can be systematically investigated by exploring individual difference factors, and the current study focuses on two factors: aptitudes for implicit learning and the amount of L2 input. First, individual differences in aptitude for implicit learning can be expected to play a role in the ultimate attainment of implicit knowledge, and in particular a stronger role in the acquisition of implicit than explicit knowledge. Research on aptitude for implicit learning is a much more recent development, compared to the long tradition of research on the role of aptitude for explicit learning, both in L2 acquisition research and in other domains, but some measures have been developed. The serial reaction time (SRT) task, in particular, has been established as a measure of aptitude for implicit learning in cognitive psychology (Kaufman et al., 2010) and found to be related to long-term attainments in L2 learning (Granena, 2013a; Linck et al., 2013). These studies support that when speakers have extensive language learning experience, the outcome of learning can be influenced by ability for implicit learning. Moreover, Granena showed a positive relationship between scores on the SRT task and knowledge of agreement structures as measured by a word monitoring task in L2 Spanish learners with many years of residence in Spain. This suggests that linguistic knowledge tapped by the word monitoring task may be acquired at least partly through the learning mechanism that was utilized in the SRT task (i.e., implicit learning) and that the word monitoring task taps into implicit knowledge. Therefore, the current study also employs the SRT task to examine how it predicts the EI and the word monitoring performance.

At the same time, it is well known that implicit knowledge takes a long time to acquire, because it "requires a very large number of encounters with each particular form" (Paradis, 2009, p. 95). Consequently, one would expect appropriate measures to draw on implicit knowledge in participants with a long length of residence (LOR) in the L2 environment, but not after less exposure. Using LOR as a proxy for the amount of L2 experience, the current study thus analyzes the results from participants with longer and shorter LOR separately.

The Current Study

The current study investigates the validity of the EI and the word monitoring tasks as measures of implicit knowledge. We focus on Japanese L2 speakers with L1 Chinese who reside in Japan, and we specifically test their linguistic knowledge of five Japanese grammatical structures involving particles. In order to examine the reconstructive processing of the EI task, spontaneous corrections of ungrammatical uses of particles are checked, and the relationship of the EI score with the working memory capacity is further examined. If the EI task is truly reconstructive, ungrammatical particles should be replaced in the repetition, and working memory should not be a strong predictor of the EI score. The primary validation of EI and word monitoring as measures of implicit knowledge is conducted by comparing the EI scores and the word monitoring performance with performance on the SRT task and the metalinguistic knowledge test. By using these two established measures for implicit (SRT task) and explicit (metalinguistic knowledge test) processes, the current study explores to what extent EI and word monitoring performance are related to those two measures, one implicit and the other explicit. The present study addressed the following research questions (RQs):

1. Is there a positive relationship between the ability to repeat grammatical structures correctly and the ability to correct ungrammatical structures in EI?

2. Is there a relationship between working memory capacity and EI scores?
3. Is there a relationship between the registration of grammatical errors measured through a word monitoring task and implicit sequence learning ability as measured with a SRT task?
4. Is there a relationship between the registration of grammatical errors and metalinguistic knowledge, as measured through a metalinguistic knowledge test?
5. Is there a relationship between EI scores and implicit sequence learning ability?
6. Is there a relationship between EI scores and metalinguistic knowledge?
7. Does LOR moderate the relationship among EI scores, the registration of grammatical errors, and implicit sequence learning ability?
8. Is there a relationship between the registration of grammatical errors and EI scores?

The first two questions were addressed to assess whether EI involved reconstructive processing. For questions three through six, we started from the premise that a sizable positive correlation between the language tasks and the SRT task (Granena, 2013a), with and little or no association with metalinguistic knowledge (R. Ellis, 2005), suggests that the language tasks tap implicit knowledge. The seventh question asked whether LOR in Japan (i.e., a proxy for the amount of experience) would change the relationship between implicit learning aptitude and the linguistic measures, because a certain amount of exposure is required in order to develop implicit knowledge (DeKeyser, 2003; Paradis, 2009). The final question aimed to investigate to what extent the registration of errors and EI performance draw on the same source of knowledge..

Method

Participants

Sixty-three L2 Japanese speakers with L1 Chinese participated in the study. They were all advanced L2 speakers, proficient enough to perform the tasks in the study. The requirement for participation in the study was advanced Japanese proficiency equivalent to N1 or N2 in the standardized Japanese Language Proficiency Test (JLPT). JLPT N1 and N2, corresponding to the previous JLPT Levels 1 and 2, are roughly equivalent to the American Council for the Teaching of Foreign Languages (ACTFL) Superior and Advanced levels, respectively, using the Oral Proficiency Interview scale (Kanno, Hasegawa, Ikeda, & Ito, 2005). JLPT Level 1 is often used as the minimum requirement for acceptance into a regular college undergraduate/graduate program in Japan.

The L2 participants were all late L2 speakers, who had arrived in Japan after the age of 18 (Table 1).[2] Their mean age was 24.65 years at the time of testing. All of them, except for one speaker, received classroom instruction before and/or after they came to Japan. Many of the participants had majored in Japanese and received four years of instruction at a university in China. Their mean LOR was 26.76 months (3–158). They were undergraduate students (n = 17), research students (n = 12), master's students (n = 27), doctoral students (n = 2), office workers (n = 2), post-doctoral or visiting scholars (n = 2), and vocational students (n = 1). Eighteen native speakers (NSs) were also recruited to serve as a baseline for performance on the EI task.

TABLE 1

Instruments

Elicited Imitation with a Built-in Word Monitoring Task

The EI procedure consists of the following three components: (a) processing of an auditory stimulus sentence; (b) a belief statement question; and (c) imitation of the sentence. In order to examine whether participants register the error during initial sentence processing, a word

monitoring task was incorporated within the EI. This new task will be referred to as the Elicited Imitation with a Built-in Word Monitoring (EIM) task. First, participants were presented with the target word in the center of the screen, and they were told to press a designated keyboard button as soon as they heard the target word in the sentence. The auditory sentence started 2 seconds after the target word appeared (long enough to read the target word for L2 speakers). The target word remained on the screen until a response was made. At the offset of the auditory sentence, the screen displayed the picture of a smiley face on the right and a sad face on the left. This prompted the participants to answer the belief statement question, "Do you agree?" They had to respond yes or no by pressing the keyboard button. After that, numbers appeared on the screen, counting down from 3 seconds to one, followed by a beep sound, accompanied by a picture of a sound speaker to cue the start of repetition. Participants were told to call out the numbers as they appeared on the screen to prevent rehearsal and rote repetition of the sentences (Mackey & Gass, 2005). Imitation of the sentence had to occur within 8 seconds. This time limit was determined based on the pilot study. Eight seconds was not too short or too long for the participants to repeat the sentence; therefore, this limit seemed to impose appropriate time pressure without too much frustration. The time limitation for repetition was imposed to make it difficult to access explicit knowledge. A bell sound indicated the end of the response time, and the next screen appeared in which participants could prepare for the next trial before hitting the space bar to proceed at their own pace.

The EIM task instructions told participants to do the following: (a) press the button as soon as they heard the target word in the sentence; (b) try to comprehend the sentence while listening, to be able to answer the belief statement question accurately; (c) upon hearing ungrammatical sentences, convert them into grammatical sentences; and (d) if necessary, use

different words in their repetition as long as those words conveyed the same meaning. The instructions told the participants to correct ungrammatical sentences if necessary, which is a different procedure from Erlam (2006), a point discussed further in the discussion section. We decided to explicitly tell participants to correct sentences so that all participants could understand the instructions in the same way. This explicit instruction prevented proficient speakers from purposely remembering and repeating grammatically incorrect sentences; this is because the pilot study with Japanese NSs showed that they tried to repeat ungrammatical sentences when the researcher did not tell them to correct these sentences. To minimize any focus on form, however, participants were not told what kind of structures they were going to be tested on,[3] or when the error would occur in the test or in which part of the sentence. Furthermore, their attention was directed to meaning, as they had to be prepared for comprehension questions.

The stimuli set consisted of 110 sentences, including 80 critical sentences containing a target structure (semantically plausible) and 30 filler sentences (semantically implausible). The EIM task needed to include no-response (not agree) sentences for the belief statement question, so implausible sentences were created as fillers (e.g., *Basukettobooru o suru toki wa, ashi de booru o takusan keru*, "When playing basketball, we kick the ball a lot"). Only plausible, critical sentences were analyzed and computed for the EI score and the word monitoring component. Plausible and implausible sentences were interspersed in the sets, and there was a short break in the middle of the task (i.e., after the 55th sentence). Before the actual test, participants were presented with eight grammatical sentences, half plausible and half implausible, to practice.

Two counter-balanced lists were created for the 80 critical sentences. The 40 grammatical sentences in List 1 had corresponding ungrammatical sentences in List 2; and the 40 ungrammatical sentences in List 1 had grammatical counterparts in List 2. Sentences were

presented in a fixed semi-random order, interspersing different types of stimulus sentences. No items testing the same structure occurred more than two times in a row. All the sentences were in the non-past tense. The sentences were comprised of vocabulary items that were familiar to the participants, so that lack of vocabulary knowledge could not interfere with the elicitation of grammatical knowledge. According to the JLPT standard (as described above), 96% of the words used for the sentences testing the five target structures (80 sentences) were within the JLPT levels (from N4 to N1) (Kokusaikouryukikin & Nihonkokusaikyouikukyoukai, 1994).[4] The remaining 4% of the words were not in the JLPT vocabulary list, but these words (mainly proper nouns) were checked by the first author to ensure that they were familiar to all participants (e.g., Fujisan, Mt. Fuji; Osaka).

A summary of sentence lengths and positions of the target words can be found in Appendix S1 in Supporting Information online. The grand mean for sentence length was 24.1 in morae and 3.1 in seconds. These indices were highly correlated to each other ($r = .83$, $p < .001$). The speech rate of the recording was 460.9 morae per minute, which approximately corresponds to the average speech rate of newscasters in modern Japanese programs, 449 morae per minute (Fukumori, 2008). The shortest recording was 2.04 seconds, and this exceeds the span of 1.5-2.0 seconds which is the time it takes for information to decay from phonological short-term memory without rehearsal or refreshing of information (Baddeley et al., 1975). The position of the target word varied from sentence to sentence (after a maximum/minimum of 4/19 morae and .06/2.7 seconds), such that participants were not able to randomly guess the position of the target word. This also means that the target structures, which occurred after the target word, were embedded in the middle of the sentences. Length and position of target words were very similar for the five structures targeted, all involving Japanese particles (described in detail below).

Metalinguistic Knowledge Test

A short metalinguistic knowledge test was administered to assess whether participants could explain the grammar rules for the five structures tested (see Appendix S2 in Supporting Information online). All target structures are usually taught explicitly in Japanese classes. The metalinguistic knowledge test included questions about whether and where participants learned the rule in each of the test items. All participants, except for two, reported that they had learned all five structures in the classroom, from grammar reference books, and/or other sources (e.g., the Internet). Since participants were all advanced speakers of Japanese, they answered the questions in Japanese. In the metalinguistic knowledge test, participants were presented with ungrammatical sentences, and the ungrammatical part was indicated with an asterisk. The grammatical alternative was given in parenthesis. Their task was to explain why the sentence was ungrammatical. One ungrammatical sentence was chosen for each of the five structures from the sentences used in the EIM task. The maximum score was five.

Working Memory Capacity: Ospan Task

The present study included a test of working memory capacity (Baddeley, 2003, 2012), because on each trial of the EI, participants listened to the sentence and retained its meaning to repeat it in the future; therefore it can be assumed that working memory ability—responsible for storage and processing of information—is involved during the EIM task. Working memory capacity was measured with the automated operation span task (Ospan) from Unsworth, Heitz, Schrock, and Engle (2005). In the Ospan task, participants first solved a math problem, indicating that the solution for the equation was correct or wrong. After each math problem, they were presented with a letter of the alphabet, and asked to remember it. After each set of math problems and letters, they were asked to select the letters in the presented order. Instructions were given in the

participants' L1 (Chinese). There were 15 sets of letters ranging from 3 to 7, and the total number of letters was 75 (3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 7).

Implicit Sequence Learning Ability: SRT Task

A probabilistic SRT task, adopted from Kaufman et al. (2010), was administered to measure aptitude for domain-general implicit sequence learning. In the SRT task, participants saw a stimulus that appeared at one of four locations on the computer screen and responded by pressing the corresponding key. Unknown to participants, the sequence of stimuli were actually generated by a probabilistic rule, and 85% of the sequences followed this rule (probable, training condition), but the other 15% of the sequence were generated by another rule (improbable, 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. The 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 is determined by the previous two locations, not just by the previous location, which makes the task more complex and implicit and minimizes chunk learning. For instance, if the first cue was 1, the probability of occurrence of the next cue (first-order conditional) is the same in Sequence A and Sequence B across the three other locations (e.g., 1–2, 1–3, 1–4). However, the probability of occurrence after the two consecutive cue (e.g., 1-2) is different between Sequence A and Sequence B. The cue is always followed by 1 in Sequence A (1-2-1), whereas it is always followed by 4 in Sequence B (1-2-4). There were eight blocks, and each block consisted of 120 trials, 960 trials in total. The SRT task was scored by subtracting the

mean RTs in the training condition from those in the control condition over blocks, which reflects the amount of learning (see Analysis section).

Target Structures

The study used five different Japanese grammar structures that were known to be difficult to acquire for Chinese speakers. Sixteen stimuli sentences, half grammatical and half ungrammatical, were created for each of the five structures (k = 80 in total). The word to be monitored was embedded immediately after the ungrammatical constructions, and the RTs to this target word were compared between the grammatical and ungrammatical items. For all the structures, the reasoning was the same: If grammatical errors were registered in real time, the RT to the monitoring word in the ungrammatical sentences would be expected to be slower than RT to the word in the grammatical sentences. The five structures are described below, with examples of both the structure and the word to be monitored.

Transitive/Intransitive Verb Pairs

Sixteen transitive/intransitive verb pairs were chosen that share the stem and morphological markings that differentiate transitive from intransitive verbs. Eight high-frequency transitive verbs were chosen (e.g., *ageru*, "raise"), whereas the other eight verbs were high-frequency intransitive verbs (e.g., *okiru*, "happen"), as shown in the following examples.

(1) *Kyuryou o/*ga ageru to, hataraku hito wa ganbaru.*

Salary-OBJECT raise if, workers-SUBJECT work harder.

If you raise the salary, workers will work harder.

(2) *Ookii jiken ga/*o okiru to, kanarazu shinbun ni deru.*

Big case-SUBJECT happens if, always newspaper-location appears.

When a big case happens, it always appears in the newspaper.

As shown in Example (1), a theme is marked by the object marking particle *o* for transitive verbs. If the subject-marking particle *ga* is used, the sentence becomes ungrammatical for the transitive verb. Example (2) demonstrates the intransitive verb usage. The subject should be marked with the subject-marking particle *ga* rather than *o*. The particles (i.e., *o* and *ga*) were manipulated rather than verb pairs because particles are less salient, and it is harder for participants to detect their ungrammaticality. The target word for monitoring is the transitive/intransitive verb (i.e., *ageru* and *okiru* in the two examples, respectively).

Wa/Ga in Relative Clauses

One of the most problematic areas for Japanese learners is *wa* and *ga*. *Wa* is a topic-marking particle, and *ga* is used to indicate the case of subject. Although Japanese speakers choose between these two particles based on meanings and contexts, the choice is sometimes determined solely by sentence structure (Minami, 1993; Noda, 1996). *Ga* is always chosen over *wa* within the relative clause. The following example illustrates the restriction. The monitoring word is a noun followed by the noun phrase (i.e., *eiga*, "movie").

(3) *Amerikajin ga/*wa tsukuru eiga wa sekaiju de mireru.*

American-SUBJECT make movie-TOPIC around the world can be seen.

The movies the Americans make can be seen around the world.

Wa/Ga in Adverbial Clauses

Similar to the restriction on the use of *wa* in relative clauses, some adverbial clauses have this constraint as well. The following adverbial clauses do not allow *wa* within the clause: hypothetical (*tara*, *ba*, *to*, *tewa*, *temo*), temporal (*toki*, *maeni*, *atode*, *made*), succession (*to*, *tara*, *te*, *renyoukei*), reason (*tame*, *te*, *kara*, *node*, *noni*), and manner (*youni*, *hodo*). It has been found that Japanese NSs do not expect another subject when the first adverbial clause contains *wa*

(Uchida, Ikegami, Ono, Oshima, & Nagatomo, 1995). A sample sentence is given in Example (4). The monitoring word is a noun at the beginning of the main clause (i.e., *gendaijin*, "modern people").

(4) *Ie ni pasokon ga/*wa nai to, gendaijin wa sukoshi fuben da to omou.*
Home computer- SUB there is not, modern people-SUB a bit inconvenient think.
If there is no computer at home, I think it is a bit inconvenient for people in the modern world.

Genitive No in Relative Clauses

The particle *no*, corresponding to English -'*s*, combines a modifying noun or clause with the following noun. A sample sentence is given in Example (5).

(5) *[Takusan Supotsu o suru] *no hito wa yaseru koto ga dekiru.*
A lot sports-OBJ play GEN person-SUB become thin can.
People [who plays sports a lot] can become thin.

The particle *no* is not allowed when a relative clause modifies a noun. However, Chinese has 的 (*de*), which is always allowed between the relative clause and the noun. Analyses of noun phrases with a relative clause in Oral Proficiency Interviews by Japanese L2 learners of Chinese showed that oversuppliance errors of *no* are prevalent between a relative clause and a modified noun (**hasitteiru no hito*, "a person that is running") even at advanced levels (Okuno, 2005; Takahashi, 2004). The target word for monitoring is the noun that is modified by the noun phrase (i.e., *hito*, "person") in Example (5).

Ni/De

Ni is used to indicate the place where a thing or a person is present, whereas *de* indicates the place where an action takes place. In other words, the particle and the state of the verb need to

agree with each other. The sample sentences in Examples (6) and (7) illustrate the differences between the two particles.

(6) *Koohi wa ie de/*ni tsukuru to kissaten yori yasui.*

Coffee-TOPIC home-LOCATION make if café than cheaper.

It is cheaper to make coffee at home than (buying one) at a cafe.

(7) *Takai apaato wa eki no chikaku ni/*de aru koto ga ooi*

Expensive apartments-TOPIC station close-to-LOCATION located it is often the case.

Expensive apartments are often located near the station.

Fourteen of the stimulus sentences were like sentence in Example (6) (*de/*ni* + action verb), whereas two sentences were similar to the sample sentence in Example (7) (*ni/*de* + stative verb).[5] The target word used for word monitoring was the verb following *ni* or *de* (i.e., *tsukuru*, "make" and *aru*, "exist").

Procedure

The participants were engaged in two one-hour individual meetings with the first author. The meetings were scheduled on two different days to reduce fatigue due to the long session. In the first meeting, the EIM task was administered. In the second meeting, the rest of the tests were administered in a fixed order: the Ospan, the SRT task, followed by the metalinguistic knowledge test. The test order was fixed to avoid differential order effects on different participants, as the current study focuses on individual differences. The most important consideration for the test order was that the most explicit task, the metalinguistic knowledge test, was administered at the end of the experiment, in order not to influence the more implicit, EIM task performance. The participants were allowed to take breaks any time they wished, but almost none needed to avail themselves of this opportunity.

Data Analysis

Elicited Imitation

First, accuracy of the belief statement questions was checked; and only those participants whose accuracy was over 85% were retained to ensure that all those included in the analysis had listened to the stimuli sentences for comprehension. None of the NSs scored less than the criterion, but two L2 speakers were excluded from further analysis. Average error rates for the belief statement questions were very low: 1.20 (*SD* = 1.54) for NSs and 4.48 (*SD* = 3.29) for L2 speakers (out of a total of 110). The first author scored the production performance based on Erlam's (2006) criteria:[6] (a) obligatory occasion created – required form supplied; (b) obligatory occasion created – required form not supplied; and (c) no obligatory occasion created. Credit was only given for the first category. The second and third categories were scored as incorrect. That means that even if the response contained an error other than the target structure, the response was scored as correct as long as the target structure was used correctly. For instance, a response such as *Okanemochi ga noru kuruma wa, kokyu no koto ga ooi* ("Rich people tend to get expensive cars") contains grammatical errors in the use of the adjective (*kokyu no* instead of *kokyu na*), but was considered correct because the target structure was used accurately (i.e., the case marker *ga* in the relative clause). Utterances self-corrected during the response time were scored correct. Giving no credit to the third category may need justification because structural modification does not necessarily mean that speakers intentionally avoid the structure. However, given that responses by NSs and L2 speakers who scored high in the EI rarely fell in the third category, it is likely that most of the responses in the third category were due to the lack of ability to use those target structures. The two sample sentences given in the instructions did not

allow any structural modifications,[7] but only word substitutions, which also means that structural modifications were not encouraged.

Word Monitoring

The RT in the word monitoring component was measured from the onset of the target word. In analyzing the RT data, outliers were discarded that were 2.5 standard deviations above or below each participant's mean, and fell outside the low and high cutoffs set at 100 and 1500 milliseconds, 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 the target word. These procedures, along with display errors, eliminated 4.3% and 8.9% of the data for NSs and L2 speakers, respectively. The dependent variable was computed as Grammatical Sensitivity Index (GSI) by subtracting the mean RT for grammatical items from the mean RT for ungrammatical items (Granena, 2013a). A higher score of the GSI means that speakers were more likely to slow down to respond to the target word in ungrammatical sentences than in grammatical sentences. The GSI thus should indicate the extent to which a participant's ability to detect errors is developed and could possibly index how robust his or her implicit knowledge is.

Metalinguistic Knowledge Test

The metalinguistic knowledge test assessed whether L2 speakers could explain why sentences were ungrammatical for each structure. Consistent with the dichotomous operationalization of metalinguistic knowledge, the responses were scored correct or incorrect, and no partial credit was given. Any explanations that were not relevant for the rule tested or were not specific enough were considered to be incorrect. For instance, most common erroneous explanations consisted of giving different usages for the *wa*/*ga* distinction that did not apply to the present

target sentences. The most commonly under-specified error was regarding the *ni*/*de* distinction (e.g., "*de* indicates location," while the correct answer was "*de* indicates location where action takes place"). The possible maximum score was five.

Ospan Task

The Ospan task was scored as the sum of all correctly recalled letters in correct positions (Unsworth et al., 2005). In other words, it gave no credit unless a set of the letters in a trial was recalled in the right positions. If an individual correctly recalled three letters in a set size of five, for example, the score was zero. In order to make sure that the Ospan task was performed appropriately, with high accuracy rates in the math problems, an 85% accuracy criterion (i.e., a maximum of 12 errors out of the 75 operations) was set for all participants, as was the case in the original study that validated this automated Ospan task (Unsworth et al.). According to this criterion, five participants were excluded. The total number of participants remaining was 56.

SRT Task

To compute the SRT scores for each participant, error responses were discarded (2.4% of trials), and outliers were identified as more than three standard deviations from the mean RT for each participant (1.7%). The amount of implicit learning was calculated by subtracting the average RT in the training condition from the average RT in the control condition from the third to the last block. The RTs in the training condition became consistently faster than those in the control condition from Block 3 to 8; the learning effect was not established in Blocks 1-2, which were excluded from the analysis (see Kaufman et al., 2010, for the same decision). Cohen's *d* used to estimate effect size across the last six blocks was .21, which was very close to .19 in Kaufman et al.'s.

Reconstructive Processing of EI

In order to examine the reconstructive processing of EI, we first computed correlations between individual L2 speakers' scores for grammatical and ungrammatical items (RQ 1). Second, the effect of working memory capacity on the EI performance was assessed by computing the correlation between the Ospan score and the EI score (RQ 2).

Validation of EI as Implicit Knowledge Measure

In order to validate EI and word monitoring as implicit knowledge measures, a set of analyses were conducted to answer RQs 3-7. First, the data were analyzed descriptively by computing correlations across test scores; this included associations between the EI score, the GSI from the word monitoring component, the SRT score, and the metalinguistic knowledge test score. On the assumption that development of implicit knowledge requires massive exposure, which should be roughly reflected in LOR, L2 speakers were divided into two groups based on the median rather than the mean of the LOR distribution as the cutoff point because the distribution was positively skewed. However, the cutoff point was adjusted, based on the median of 20 months, by adding 10 additional months to secure a sufficient number of participants in both groups: the long LOR group (LOR >= 30 months, n =19) and the short LOR group (LOR < 30 months, n =42). Although this LOR cutoff point may seem arbitrary, a subsequent multivariate analysis of covariance (MANCOVA) included LOR as a continuous covariate to determine the role of LOR directly.

A MANCOVA was conducted to examine the interactions between LOR, SRT score, and metalinguistic knowledge, with the EI score and the GSI used as the dependent variables. Such an analysis can reduce redundant variances among the covariates and the dependent variables, estimating their effects on the dependent variables (i.e., the EI score and the GSI) more accurately, while providing better control for Type I errors, compared to conducting separate

analyses of covariance (ANCOVAs) (Tabachnick & Fidell, 2007). Therefore, a MANCOVA was conducted using the total EI score and the total GSI as the dependent variables, with high/low SRT grouping based on the median score as a fixed factor, and LOR and metalinguistic knowledge as covariates. Interaction terms (SRT×LOR, Metalinguistic Knowledge×LOR, and SRT×Metalinguistic Knowledge) were added, in addition to the group and covariate terms. The alpha level was set at .05; and effect sizes were reported (partial η^2), using the following criteria: small (0.01), medium (0.06), and large (0.14) (Cohen, 1988). The last RQ was addressed by computing the correlation between the EI score and the GSI from the word monitoring.

Results

Full descriptive statistics and reliability indices for all tests are reported in Appendix S3 as part of Supporting Information online.

Elicited Imitation

NSs' imitation performance was nearly perfect (*M* = 78.61 out of a total of 80, *SD* = 1.09), which suggests that the test items were appropriate for high-level users of Japanese. The NSs' mean scores were 39.61 (*SD* = 0.70) and 39.00 (*SD* = 1.08) for grammatical and ungrammatical items, respectively. The total EI score for L2 speakers was 58.11 (*SD* = 14.18), and the distribution was normal by a Kolmogorov-Smirnov (K-S) test ($p > .05$). The mean scores were 30.26 (*SD* = 7.41) and 27.85 (*SD* = 7.23) separately for grammatical and ungrammatical items. Internal consistency indexed by Cronbach's alpha was very high for L2 speakers (List 1= .93, List 2 = .95).

Word Monitoring

In the NS group, the mean RT for the grammatical sentences was 394 ms (*SD* = 71) and that for the ungrammatical sentences was 495 ms (*SD* = 107). A paired-samples *t*-test was performed to compare the difference in RTs in order to show that the task was designed well and appropriately

performed by the NSs. A significant difference was detected with a large effect size, $t(17) = 7.81$, $p < .001$, Cohen's $d = 1.14$. In the L2 group, the mean RT for the grammatical sentence was 539 ms ($SD = 103$) and that for the ungrammatical sentences was 555 ms ($SD = 101$). A paired-samples *t*-test showed a significant difference with a small effect size, $t(60) = 2.33$, $p = .023$, $d = .16$. The RT difference (GSI) was computed as follows: mean ungrammatical RT – mean grammatical RT. The total GSIs were much larger for the NSs ($M = 101$ ms, $SD = 52$) than for the L2 speakers ($M = 16$ ms, $SD = 54$). The GSI was normally distributed in the L2 speaker group, according to the K-S test ($p > .05$). Split-half reliability with the Spearman-Brown correction was high for the L2 speakers (List 1= .91, List 2 = .86).

Metalinguistic Knowledge Test

The mean score for the metalinguistic knowledge test was 3.32 out of 5 ($SD = 1.59$), which means that the L2 speakers were able to explain the rules for three structures on average. Examining individual scores for each target structure shows that most structures were known by the L2 speakers (see Table 2). More than 70% of the participants were able to explain the rules about transitive/intransitive verb pairs, the genitive particle, and the location particle. For the usages of *wa* and *ga*, about half of the participants knew the relevant grammatical rules. The distribution of the score was not normal according to the K-S test ($p < .001$). Cronbach's alpha for this test was .76.

TABLE 2

Ospan Task

The average number of errors in math operations was 5.61 ($SD = 2.77$) out of 75. The mean score for the memory task was 49.73 ($SD = 14.73$), and the distribution was normal according to the K-S test ($p > .1$). Cronbach's alpha was .69.

SRT Task

The mean score for the SRT task was 17 ms (*SD* = 13). According to the K-S test, the distribution of the SRT scores was normal ($p > .05$). Split-half reliability, with the Spearman-Brown correction, was .42. In light of other studies of implicit learning in both psychology and L2 acquisition, the reliability index is deemed acceptable above .4 (Granena, 2013a; Kaufman et al., 2010; Reber, Walkenfeld, & Hernstadt, 1991). While a lower reliability of a measure usually attenuates possible observed correlations with other variables due to measurement error, previous studies that used the SRT task found significant correlations with scores on verbal analogical reasoning and processing speed (Kaufman et al., 2010) or acquisition of agreement structure in L2 (Granena, 2013a). If in the current study a significant relationship of the SRT task with other tasks was found, despite its lower reliability, this strengthens evidence for the relationship.

Reconstructive Processing of EI

Reconstructive processing of the EI was examined through comparisons of performance on the grammatical and ungrammatical items. A strong positive correlation was detected ($r = .88$, $p < .001$), which provides evidence for spontaneous correction of ungrammatical sentences or reconstructive processing. Another criterion for reconstructive processing was tested by examining the role of working memory capacity in the EI performance. There was a weak positive relationship between the total EI scores and the Ospan scores ($r = .29$, $p = .036$). The correlation was significant for grammatical items ($r = .31$, $p = .024$), but not for ungrammatical items ($r = .24$, $p = .077$).

Relationship of EI and Word Monitoring to Implicit Learning Aptitude and Metalinguistic Knowledge

In order to investigate whether the word monitoring performance (GSI) and the EI performance tap implicit knowledge, the correlations of each of these two scores with the measure of implicit learning ability (SRT) were calculated. As implicit knowledge can be expected to result largely from implicit learning, implicit learning is predicted by the SRT task. Neither of the correlation coefficients was meaningful or significant (r = .04, p = .780 for GSI and r = .08, p = .554 for the EI) for the participant sample as a whole. The correlations between the GSIs (from the word monitoring component) and the SRT scores were computed again, separately for the two LOR groups (see Figure 1). Strikingly, a moderate positive relationship emerged only in the long LOR group (r = .43, p = .065). In contrast, the relationship between GSIs and SRT scores for short LOR speakers was small and negative (r = –.19, p = .218).

FIGURE 1

The same analysis was performed using the EI scores for the short and long LOR groups. We expected to observe a positive relationship for the long LOR group; however, there was no relationship between the EI scores and the SRT scores in the long LOR group (r = .02, p = .928) or in the short LOR group (r = .13, p = .421). A scatter plot of these data with a regression line indicates no relationship between the total EI scores and SRT in either of the LOR groups (Figure 2). These divergent results for the GSIs and EI scores suggest that word monitoring draws on knowledge that was acquired partly through implicit learning mechanisms, while EI might be drawing on different sources of knowledge.

FIGURE 2

Given that word monitoring and EI may draw on different sources of knowledge, correlations of metalinguistic knowledge with both the GSIs and the EI scores were computed to probe the difference further. No significant relationship of metalinguistic knowledge was

detected with the GSI for the whole group (r = .15, p = .239), for the long LOR group (r = .06, p = .815), or for the short LOR group (r = .17, p = .281). Figure 3 graphically shows that no significant relationship existed between the total GSI and metalinguistic knowledge for either LOR group.

FIGURE 3

Overall, the EI score, in contrast, was significantly related to metalinguistic knowledge in the whole group (r = .46, p < .001). As shown in Figure 4, metalinguistic knowledge was related positively with the EI performance especially in the long LOR group (r = .70, p = .001), whereas the magnitude of the correlation coefficient was smaller in the short LOR group (r = .33, p = .032). In sum, metalinguistic knowledge scores were not correlated with the word monitoring performance, while they were significantly correlated with the EI scores, to varying degrees depending on LOR.

FIGURE 4

The preceding analyses have revealed interactions between LOR, the SRT score, and metalinguistic knowledge: the EI score was found to be significantly correlated with metalinguistic knowledge, but not with the SRT task. The GSI, on the contrary, was not correlated with metalinguistic knowledge, but positively related with the SRT score, and only for the long LOR group. These interactions were further examined using a MANCOVA.[8] The assumptions of equality of error variances (Levene's test) and the equality of covariances (Box's test) were met for the data (p = .18). At the multivariate level, the MANCOVA revealed a main effect of metalinguistic knowledge which was approaching significance, with a medium effect size, $F(2, 53) = 2.32$, p = .108, partial η^2 = .08, as well as a significant SRT × LOR interaction, also with a medium effect size, $F(2, 53) = 2.47$, p = .094, partial η^2 = .09. At the univariate level,

Metalinguistic Knowledge was a significant covariate for EI, with a medium effect size, $F(1, 54) = 4.66$, $p = .035$, partial $\eta^2 = .080$. With the GSI as the dependent variable at the univariate level, a marginally significant SRT × LOR interaction was also found, with a medium effect size, $F(1, 54) = 3.034$, $p = .087$, partial $\eta^2 = .053$. These results were consistent with the correlation patterns, and further corroborated that metalinguistic knowledge played a significant role in the EI performance, and that the effect of SRT on the GSI was moderated by LOR.

Relationship Between EI and Word Monitoring

Finally, the relationship between the GSIs and the EI scores was investigated. The overall correlation was significant but modest ($r = .37$, $p = .003$), and the correlation coefficients were almost identical for grammatical ($r = .357$, $p = .005$) and ungrammatical sentences ($r = .359$, $p = .004$). Since the SRT scores were correlated exclusively with the GSIs for the long LOR speakers, we expected to observe no significant relationship between GSIs and EI scores for this group. The correlations were computed separately for the short and long LOR groups. As predicted, the correlation coefficient between GSI and EI in the long LOR group was lower ($r = .25$, $p = .299$) than that in the short LOR group ($r = .44$, $p = .004$). The correlation coefficients for grammatical and ungrammatical sentences did not differ at all ($r = .26$, $p = .287$ and $r = .24$, $p = .332$ for the long LOR group, and $r = .41$, $p = .007$ and $r = .43$, $p = 004$ for the short LOR group, respectively).

Discussion

Reconstructive Processing in Elicited Imitation

The first assumption for the reconstructive EI was a positive relationship between the EI scores for grammatical and ungrammatical sentences. A strong positive relationship between the grammatical and ungrammatical items was found ($r = .88$, $p < .001$), which is even stronger than

the coefficient (r = .73) obtained in the previous study by Erlam (2006). L2 speakers who can correct the ungrammatical sentences in the EI tend to repeat the grammatical sentences more accurately, which suggests that the ability to reconstruct the sentence using linguistic knowledge is required for both types of sentences. This is clear evidence for reconstructive processing in the EI. We also examined the effect of working memory capacity on the performance of EI directly. The Ospan score was only weakly correlated with the EI score (r = .29, p = .036). When the EI score was broken down into grammatical and ungrammatical sentences, only the EI score for the grammatical sentences was significantly correlated with working memory (r = .31, p = .024). In sum, the current study provided evidence showing that EI performance involves reconstructive processing, and working memory capacity may influence the EI performance, particularly for grammatical sentences, but the effects of working memory appear to be limited.

Elicited Imitation and Word Monitoring Performance

The primary goal of this study was validating EI and word monitoring tasks as measures of implicit knowledge. Implicit learning aptitude, as measured through the SRT task, was predicted to correlate with scores on a test which draws on implicit linguistic knowledge. When all L2 speakers with short and long LORs were included, there was no relationship between the SRT scores and either the EI scores or the GSIs from the word monitoring component. An intriguing interaction between the SRT scores and LOR was, however, revealed: The aptitude for implicit learning was related to the GSI, not the EI score, and only among the long LOR speakers. This suggests that only L2 speakers with sufficient L2 exposure can draw on implicit knowledge in word monitoring, and that implicit knowledge was reliably assessed through the word monitoring component, but not through the EI. This is further supported by the weak relationship

between the word monitoring and EI performance, particularly in the long LOR group (r = .25, p = .299).

Metalinguistic knowledge, however, was a significant predictor of EI (with a medium effect size), but did not play a role in the word monitoring performance, suggesting that EI draws on explicit types of linguistic knowledge. Even though EI involved reconstructive processing, and time pressure was imposed for imitation, it appears that participants were able to use their explicit knowledge before and/or while they imitated the sentence. This means that L2 speakers seem to have used linguistic knowledge with awareness while performing the EI. Given the time pressure during the task, explicit knowledge had to be accessed quickly to respond in the EI; it is reasonable to assume, therefore, that the construct which EI measures is automatized explicit knowledge, not implicit knowledge. Furthermore, the relationship between the EI performance and metalinguistic knowledge seems stronger in the long LOR group than in the short LOR group, although a significant interaction between metalinguistic knowledge and LOR was not found. It appears, then, that L2 speakers with less L2 exposure are less adept at utilizing metalinguistic knowledge when performing the EI under time pressure. In other words, L2 speakers in the short LOR group did not acquire explicit knowledge that is deployed quickly (automatically) in EI, whereas L2 speakers in the long LOR group possessed more automatized explicit knowledge to deploy.

The present study's findings run counter to the claim, proposed in the L2 acquisition literature, that EI is one of the best measures of implicit knowledge (R. Ellis et al., 2009). R. Ellis demonstrated that three implicit knowledge measures (i.e., oral narrative, timed GJT, and EI) and two explicit knowledge measures were loaded onto separate factors; this finding was replicated in the subsequent studies (Bowles, 2011; Zhang, 2014). The critical factor that differentiated the

implicit and explicit types of measures was the presence or absence of time pressure, which has also been found in other studies to influence what type of knowledge is tapped (Erlam, 2006; Granena, 2013i; Gutiérrez, 2013; Han & Ellis, 1998; Loewen, 2009). Time pressure does not, however, guarantee the inaccessibility of explicit knowledge, particularly when explicit knowledge has been automatized (DeKeyser, 2003, 2009). In fact, the two constructs extracted from the factor analysis in R. Ellis' study could be labeled automatized explicit knowledge and less automatized explicit knowledge. Without independent measures, for example, of implicit/explicit learning aptitude, it is difficult to determine whether the less explicit factor in the analyses such as those reported by R. Ellis and colleagues represents implicit knowledge or automatized explicit knowledge.

Limitations

The current study provides the first empirical evidence that challenges the validity of the EI task. While this study has several limitations, it opens up new avenues for future research. First, the procedure used for EI in the current study was different from that employed in previous studies, which limits the generalizability of our findings to EIs with different procedures. The word monitoring component in the current EI led the participants to focus on listening for a lexical item, and this additional component might have drawn their attention more to formal aspects of the stimulus sentences, compared to their meaning. Additionally, some monitoring words were part of the target grammatical structures. Particularly, in the transitive/intransitive sentences and the *ni*/*de* locatives, the target word was a transitive or an intransitive verb, which might have drawn participants' attention to the verb form. Further research could administer an EI task and a word monitoring task separately to investigate whether it can replicate the current findings.

Second, since half of the participants had majored in Japanese linguistics or Japanese education, this specific background of the L2 speakers in the present study made EI conducive to the retrieval of automatized explicit knowledge. Different L2 populations, such as heritage speakers or naturalistic L2 learners with less instructional experience, may rely on implicit knowledge to a greater extent (e.g., Bowles, 2011). Third, in order to avoid a situation in which speakers would think that they were supposed to repeat ungrammatical sentences exactly as they heard them, the EI instructions in the present study required participants to correct the errors. This might have drawn more of the participants' attention to form than in Erlam's (2006) study, where the instructions were to "repeat the sentences in *correct* English" (emphasis added). In the practice session before the main experiment, eight statements, half grammatical and half ungrammatical, were provided to show how ungrammatical sentences were converted to grammatical sentences. Since Erlam's participants were not explicitly told that ungrammatical sentences should be converted to grammatical sentences, instructions in Erlam's study could be more implicit than those in the current study. The instructions in the present study to correct ungrammatical sentences, albeit unavoidable, might have pushed the L2 speakers to use explicit knowledge more than in previous studies. Possibly, using less salient grammatical errors might obviate the need for instructions to correct ungrammatical sentences.

Fourth, another factor that needs to be addressed in future research is the time limit for EI. The current study imposed an 8 second time limit per sentence, but it might not have put enough time pressure on all the participants in the current study. Fifth, the metalinguistic knowledge test included only one sentence for each target structure because metalinguistic ability can be measured dichotomously with this format (Granena, 2013a). It may be possible that some people might respond differently to different sentences with the same target structures, and more test

items should have been included in the metalinguistic knowledge test in order to estimate participants' ability more stably. Further research should address this issue with the metalinguistic knowledge (cf. R. Ellis, 2005). In addition, the present study assessed metalinguistic knowledge, but did not measure explicit learning aptitude. With explicit learning aptitude tests, a more adequate comparison could be made between the effects of implicit and explicit learning aptitude on the acquisition of explicit and implicit knowledge. Finally, the conclusions about the associations between the word monitoring component and implicit learning aptitude would be strengthened if future research included both subjective (e.g., interviews) and objective (e.g., recognition tests) measures to check the lack of explicit knowledge that could develop through the SRT task (Reed & Johnson, 1994).

Conclusion and Suggestions for Further Research

In the line of research on the validity of EI as a measure of implicit knowledge (Bowles, 2011; R. Ellis, 2005; Erlam, 2006; Zhang, 2014), the present study set out to examine whether EI is a reliable measure for implicit knowledge by shedding light on the internal processing of the auditory stimulus. Results showed that EI involves reconstructive processing with only marginal effects from individuals' working memory capacity, which satisfies the requirement that EI involve reconstructive processing. The present findings, however, demonstrated that word monitoring performance can reliably reflect the use of implicit knowledge when L2 speakers have access to implicit knowledge, whereas EI appears to be a less valid measure for implicit knowledge. Even with time pressure imposed on the imitation, L2 speakers can still access automatized explicit knowledge. This led us to claim that EI, while certainly better than untimed GJTs and possibly a good alternative to timed GJTs or oral narratives depending on the research purposes, is nevertheless too coarse a measure for implicit knowledge, which cannot completely

shut off access to automatized explicit knowledge. In contrast, the word monitoring task seems a more fine-grained measure for implicit knowledge, and as argued at the outset of this study, it can assess language users' online use of linguistic knowledge with little or no awareness while their attention is directed to meaning. Given difficulties with operationalizing and measuring awareness, which is usually accomplished through participant self-reports (but see Rebuschat, 2013), it may be possible to formulate a more parsimonious criterion for implicit knowledge than through direct self-assessment of awareness. Such a criterion might involve registration or detection (in the restricted sense) of errors during real-time sentence processing for comprehension. With this new criterion, experimental psycholinguistic methods are expected to play a more important role in the measurement of explicit and implicit knowledge in the future.

Final revised version accepted 18 December 2014

Notes

1 Three of those criteria (systematicity, certainty, and learnability) have been argued to provide evidence that implicit knowledge is being measured, but their validity can be questioned, and these criteria were not supported in Ellis' (2005) own large-scale study (but see Gutiérrez, 2013, for evidence of systematicity).

2 One participant had lived in Japan between the ages of 11 and 13. She received classroom instruction only when she was 20, however, and therefore was not considered to be different from other learners.

3 In the instructions, they were given only one example of an ungrammatical sentence that was not tested (i.e., *Beatles wa yumei katta* can be rephrased as *Beatles wa yumei datta*, "The Beatles were famous"). The ungrammatical part of this sentence was the conjugation of the adjective,

yumei na (“famous”). The Japanese adjectives have past-tense inflections, and *yumei na* should be conjugated to *yumei datta* in the past tense.

4 We used the Reading Tutor to check vocabulary levels (http://language.tiu.ac.jp/index_e.html).

5 The particle *ni* takes a limited number of stative verbs (e.g., *iru*, “exist”; *sumu*, “live”).

6 The rater rarely found cases that were difficult to score, as the recordings were very clear.

7 For example, the sentence had modifications in the lexical items, not the sentence structure. A sentence such as *totemo hansamu na dansei wa se ga takai koto ga ooi* was converted to *kakkoii otokono hito wa totemo se ga takai koto ga ooi* (“Handsome boys tend to be tall”).

8 Since the LOR and the metalinguistic knowledge test scores were not normally distributed, those scores were transformed to reduce skewness. Because the MANCOVA results with the transformed variables showed essentially the same patterns from the non-transformed data, only the results of the analyses using non-transformed data are reported.

References

Baddeley, A. D. (2003). Working memory and language: an overview. *Journal of communication disorders, 36*(3), 189-208. doi: 10.1016/S0021-9924(03)00019-4

Baddeley, A. D. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology, 63*, 1-29. doi: 10.1146/annurev-psych-120710-100422

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. doi: 10.1016/S0022-5371(75)80045-4

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). Hillside, NJ: Lawrence Erlbaum.

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge. *Studies in second language acquisition, 33*(2), 247-271. doi: 10.1017/S0272263110000756

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. doi: 10.1177/0267658313493432

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coughlin, C. E., & Tremblay, A. (2013). Proficiency and working memory based explanations for nonnative speakers' sensitivity to agreement in sentence processing. *Applied Psycholinguistics, 34*(3), 615-646. doi: 10.1017/S0142716411000890

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. (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.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in second language acquisition, 27*(2), 305-352. doi: 10.1017/S027226310505014X

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in second language acquisition, 27*(2), 141-172. doi: 10.1017/S0272263105050096

Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J., & Reinders, H. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*. Tonawanda, NY: Multilingual Matters.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied linguistics, 27*(3), 464-491. doi: 10.1093/applin/aml001

Fraser, C., Bellugi, U., & Brown, R. (1963). Control of grammar in imitation, comprehension, and production. *Journal of verbal learning and verbal behavior, 2*(2), 121-135. doi: 10.1016/S0022-5371(63)80076-6

Fukumori, T. (2008). The speech rate of announcers and news-casters: Based on a news program boradcasted on May 3, 2006. *Daito Bunka Daigaku Gaikokugogakubu Soritsu Sanjuugoshuunen kinenshu* (pp. 191-209). Shinjuku-ku, Tokyo: Morimoto Insatsu.

Granena, G. (2013a). Individual Differences in Sequence Learning Ability and Second Language Acquisition in Early Childhood and Adulthood. *Language Learning*, *63*(4), 665–703. doi: 10.1111/lang.12018

Granena, G. (2013i). 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.

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. doi: 10.1017/S0272263113000041

Hama, M., & Leow, R. P. (2010). Learning without Awareness Revisited. *Studies in second language acquisition*, *32*(3), 465-491. doi: 10.1017/s0272263110000045

Han, Y., & Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research*, *2*(1), 1-23. doi: 10.1177/136216889800200102

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. doi: 10.1017/S0272263105050084

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. doi: http://dx.doi.org/10.3138/cmlr.64.1.215

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. doi: 10.1111/j.1467-9922.2010.00627.x

Kanno, K., Hasegawa, T., Ikeda, K., & Ito, Y. (2005). Linguistic Profiles of Heritage Bilingual Learners of Japanese. In J. Cohen, K. T. McAlister, K. Rolstad, & J. MacSwan (Eds.), *Proceedings of the 4th International Symposium on Bilingualism* (pp. 1139-1151). Somerville, MA: Cascadilla Press.

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. doi: 10.1016/j.cognition.2010.05.011

Kokusaikouryukikin, & Nihonkokusaikyouikukyoukai. (1994). *Japanese Language Proficiency Test: Test Content Specifications*. Tokyo, Japan: Bonjinsha.

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

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. doi: 10.1111/lang.12011

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.

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

Minami, F. (1993). *Gendai nihongo bunpou no rinkaku [The Structure of Modern Japanese]*. Bunkyo-ku, Tokyo: Taishuukanshoten.

Noda, H. (1996). *New Series of Japanese Grammar: Wa and Ga*. Bunkyo-ku, Tokyo: Kuroshio Shuppan.

Okuno, Y. (2005). *Dainigengo to shite no nihongo shutoku katei ni okeru gengoteni no kenkyu: no no kajou shiyou wo chushin to shite [Linguistic Transfer in Acquisition of Japanese as a Second Language: On the Over-suppliance Usage of No]*. (Unpublished doctoral dissertation), Hiroshima University.

Okura, E., & Lonsdale, D. (2012). Working memory's meager involvement in sentence repetition tests. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 2132-2137). Austin, TX: Cognitive Science Society.

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

Posner, M. (1994). Attention in congnitive neuroscience: An overview. In MS Gazzaniga. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences.* (pp. 615-624). Cambridge, MA: MIT Press.

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. doi: http://dx.doi.org/10.1037/0278-7393.17.5.888

Rebuschat, P. (2013). Measuring Implicit and Explicit Knowledge in Second Language Research. *Language Learning, 63*(3), 595–626. doi: 10.1111/lang.12010

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. doi: http://dx.doi.org/10.1037/0278-7393.20.3.585

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. doi: 10.1177/0267658313503171

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

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Harper & Row.

Takahashi, O. (2004). Rentaishushoku kouzou no shutoku ni kansuru kenkyu gaikan: 'No' no kajoushiyou to datsuraku wo chushin ni [The acquisition of noun modifying structures in Japanese : The overuse and omission of the particle "no"]. *Gengo Bunka to Nihongo Kyoiku*, 147-167.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in second language acquisition, 16*, 183-203. doi: 10.1017/S0272263100012870

Uchida, A., Ikegami, M., Ono, S., Oshima, Y., & Nagatomo, K. (1995). Yosokubunpou kenkyu (1): 'ga' to 'wa' no yosokukinou ni tsuite. [A Study on Expectancy Grammar (1): On the Function of 'Ga' and 'Wa']. *Gengo Bunka to Nihongo Kyoiku (Mizutani Nobuko Sensei Taikan Kinengo), 9*, 134-159.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498-505. doi: 10.3758/BF03192720

Vinther, T. (2002). Elicited imitation: a brief overview. *International Journal of Applied Linguistics, 12*(1), 54-73. doi: 10.1111/1473-4192.00024

Williams, J. N. (2005). Learning without awareness. *Studies in second language acquisition, 27*(2), 269-304. doi: 10.1017/S0272263105050138

Williams, J. N. (2009). Implicit learning in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 319-353). Bingley, UK: Emerald Group Publishing.

Zhang, R. (2014). Measuring University-Level L2 Learners' Implicit and Explicit Linguistic Knowledge. *Studies in second language acquisition, First View*, 1-30. doi: http://dx.doi.org/10.1017/S0272263114000370

Table 1 Background information for L2 speakers

Variable	*M*	*SD*	*Min*	*Max*
Age (years)	24.65	4.34	20	52
Starting age of instruction (months)	19.13	2.20	14	25
Length of instruction (months)	39.03	20.17	0	125
Age of arrival (months)	22.17	3.16	18	42
Length of residence (months)	26.76	26.82	3	158

Table 2 The number of correct answers (percentage) in the metalinguistic knowledge test by target structure

Structure	Correct descriptions (n = 63)
Transitive/intransitive	51 (81%)
Wa/*ga* in adverbial clauses	26 (41%)
Wa/*ga* in relative clauses	34 (54%)
Genitive	53 (84%)
Ni/*De*	45 (71%)

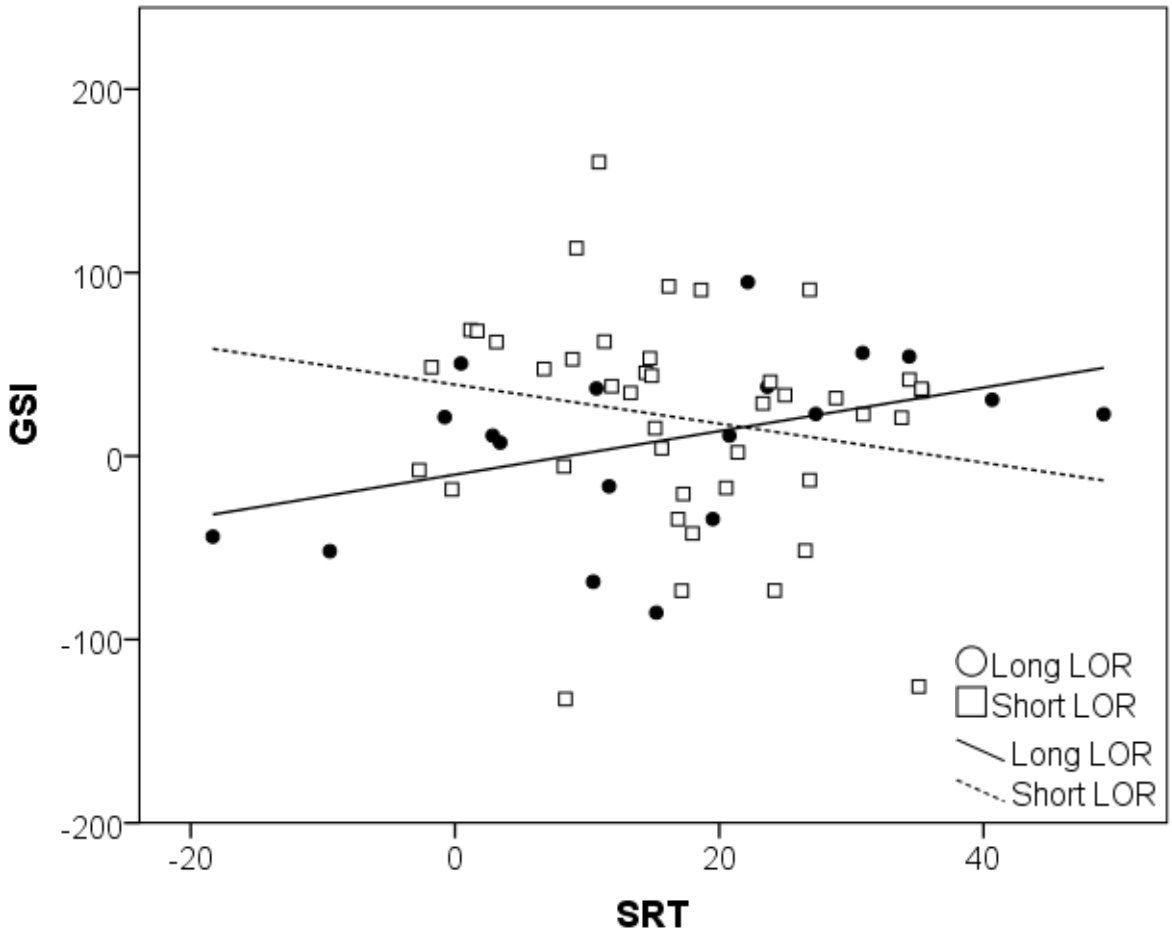

Figure 1 Relationship between GSI and SRT for long and short LOR groups.

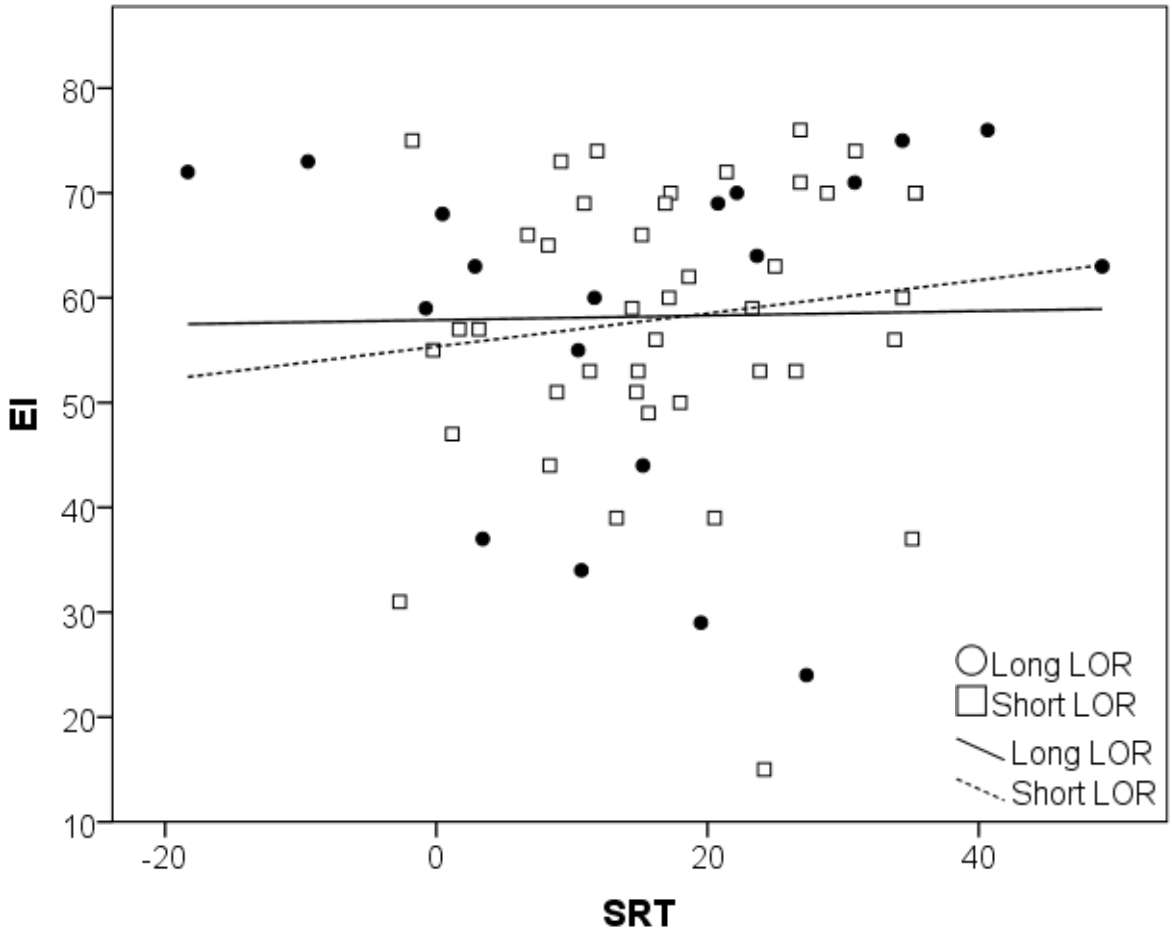

Figure 2 Relationship between EI and SRT for long and short LOR groups.

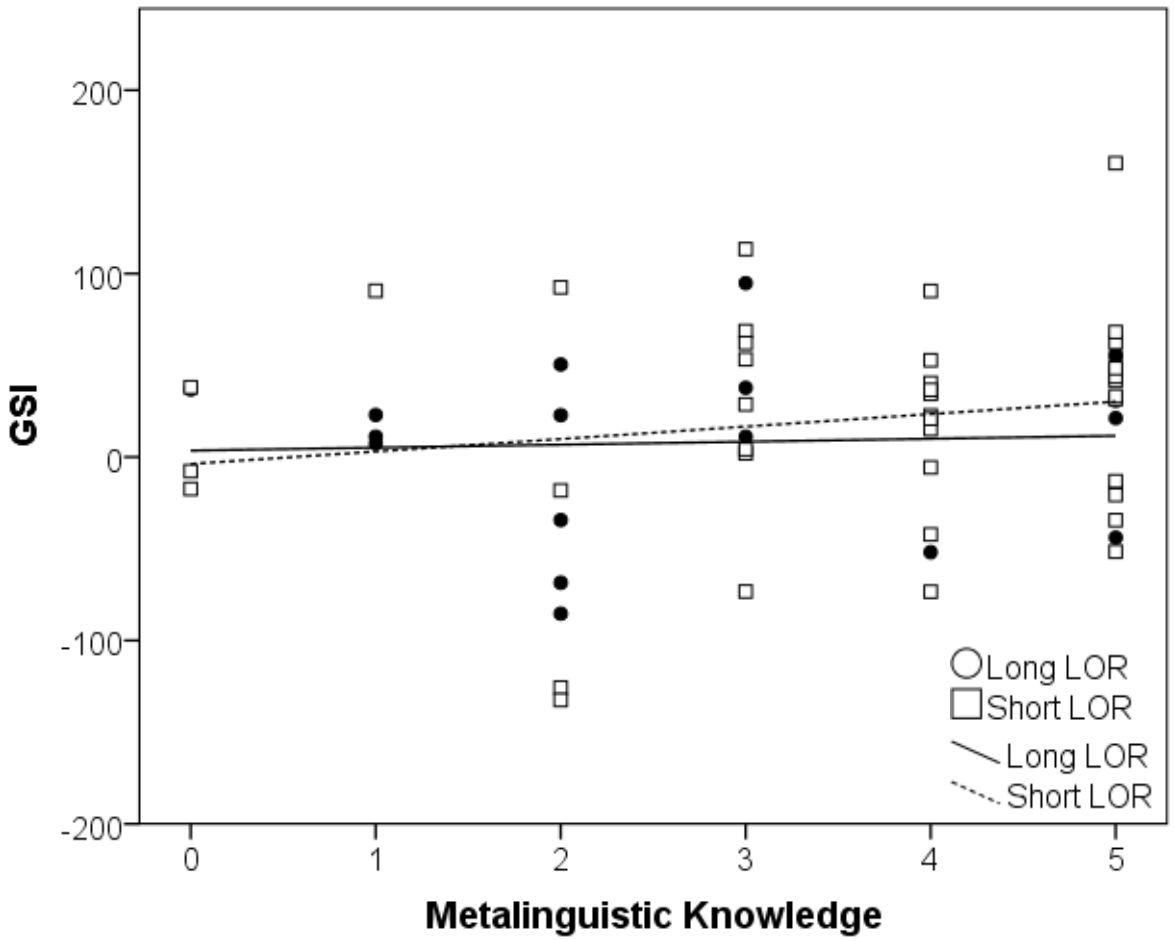

Figure 3 Relationship between GSI and metalinguistic knowledge.

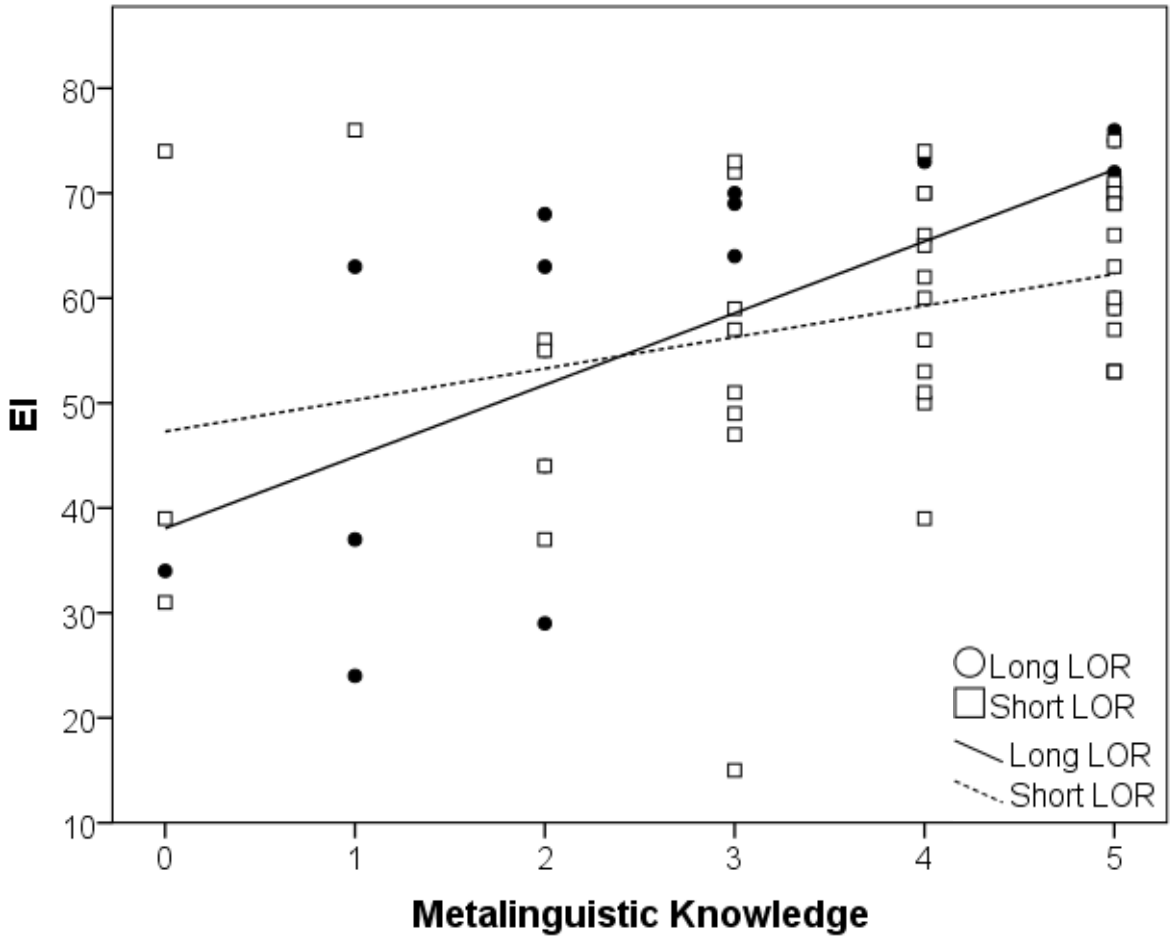

Figure 4 Relationship between EI and metalinguistic knowledge.