Applied Psycholinguistics, page 1 of 33, 2017
doi:10.1017/S014271641700011X

Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge

YUICHI SUZUKI
Kanagawa University

Received: July 28, 2016 Accepted for publication: March 27, 2017

ADDRESS FOR CORRESPONDENCE
Yuichi Suzuki, Department of Cross-Cultural Studies, Kanagawa University, 3-27-1, Rokkakubashi, Kanagawa-ku, Yokohama-shi, Kanagawa, 221–8686, Japan. E-mail: szky819@kanagawa-u.ac.jp

ABSTRACT
Accumulating evidence suggests that time-pressured form-focused tasks like grammaticality judgment tests (GJTs) can measure second language (L2) implicit knowledge. The current paper, however, proposes that these tasks draw on automatized explicit knowledge. A battery of six grammar tests was designed to distinguish automatized explicit knowledge and implicit knowledge. While three time-pressured form-focused tasks (an auditory GJT, a visual GJT, and a fill in the blank test) were hypothesized to measure automatized explicit knowledge, three real-time comprehension tasks (a visual-world task, a word-monitoring task, and a self-paced reading task) were hypothesized to measure implicit knowledge. One hundred advanced L2 Japanese learners with first language Chinese residing in Japan took all six tests. Confirmatory factor analysis and multitrait-multimethod analysis provided an array of evidence supporting that these tests assessed two types of linguistic knowledge separately with little influence from the method effects. The results analyzed separately by length of residence in Japan (a proxy for the amount of naturalistic L2 exposure) showed that learners with longer residence in Japan can draw on implicit knowledge in the real-time comprehension tasks with more stability than those with shorter residence. These findings indicate the potential of finely tuned real-time comprehension tasks as measures of implicit knowledge.

The issue of implicit and explicit knowledge and learning mechanisms has attracted attention from many second language acquisition (SLA) researchers because of its theoretical and educational implications (e.g., Hulstijn, 2005). To tackle the issues surrounding explicit and implicit knowledge and learning (e.g., interface issues), the methodological problem of measuring implicit knowledge is crucial (Suzuki & DeKeyser, 2017). Explicit and implicit knowledge are distinguished based on awareness; implicit knowledge is deployed without awareness, whereas explicit knowledge requires some level of awareness (DeKeyser, 2003; Williams, 2009). Previous SLA studies have shown empirically that explicit and implicit knowledge are distinct constructs that can be measured separately (Bowles, 2011; R. Ellis, 2005; Gutiérrez, 2013; Zhang, 2015). A recent study, however, employed a more finely tuned psycholinguistic technique to examine real-time grammar processing and cast doubt on the validity of existing implicit knowledge measures (Suzuki & DeKeyser, 2015; Vafaee, Suzuki, & Kachinske, 2017).

The present paper reports a construct validation study of a new battery of finely tuned tests for implicit knowledge: the eye-tracking-while-listening task, the word-monitoring task, and the self-paced reading task. These real-time comprehension tasks were compared with the existing tasks that have been claimed to measure implicit knowledge, for example, timed grammaticality judgment tests (GJT), which were hypothesized to draw more on automatized explicit knowledge in this study.

PROBLEMS IN PREVIOUS MEASURES OF "IMPLICIT" KNOWLEDGE IN SLA

A seminal study by R. Ellis (2005, 2009) developed three tests that were hypothesized to measure implicit knowledge: an oral narrative task, a timed GJT, and an elicited imitation (EI) task. Since these tasks were performed under time pressure, Ellis claimed that second language (L2) learners are more likely to draw on implicit knowledge. He conducted a confirmatory factor analysis (CFA) and demonstrated that these time-pressured tests were loaded onto a separate factor from an explicit knowledge factor that untimed tests were loaded onto (i.e., an untimed GJT and a metalinguistic knowledge test). This finding was essentially replicated in subsequent studies with different L2 populations (Ercetin & Alptekin, 2013; Gutiérrez, 2013; Sarandi, 2015; Zhang, 2015) and a heritage learner population (Bowles, 2011).

The critical methodological factor that differentiated implicit knowledge from "explicit" knowledge in those studies was imposing time pressure on the language tests; however, time pressure cannot necessarily limit access to explicit knowledge enough to ensure that implicit knowledge is drawn upon (DeKeyser, 2003; Suzuki & DeKeyser, 2015; Vafaee et al., 2017). Proficient L2 learners may still access explicit knowledge with awareness even if the execution is rapid (i.e., automatized explicit knowledge), which is distinguished from the use of linguistic knowledge *without* awareness (i.e., implicit knowledge). In other words, both implicit knowledge and automatized explicit knowledge are accessed quickly, but they are distinguished based on the awareness criterion. Highly automatized knowledge is conscious knowledge that one can draw on quickly. It is functionally equivalent to implicit knowledge in the sense that it is not easy to distinguish behaviorally (and impossible to distinguish in mundane language use), but it remains knowledge that one is aware of, and awareness is the defining criterion of explicit knowledge. In cognitive psychology, automaticity (i.e., the end point of automatization) is often characterized as lack of awareness (e.g., Jacoby, 1991; Posner & Snyder, 1975). However, automatization is a long process, and even highly automatized skills do not always become 100% automatic, particularly with complex skills like L2 learning. Automatization of explicit knowledge should be regarded as a gradual development, not an all or nothing phenomenon (DeKeyser, 2015). In the current paper, automatized explicit knowledge is thus defined as a body of conscious linguistic knowledge including different levels of automatization. An attempt is made to measure *partially* (not fully) automatized linguistic knowledge with a conscious correlate, which can be theoretically distinguishable from implicit knowledge.

A recent study provided evidence that it is possible to devise linguistic tasks that can draw upon implicit knowledge separately from automatized explicit knowledge (Suzuki & DeKeyser, 2015). Suzuki and DeKeyser (2015) demonstrated that the EI task, which was the best measure of implicit knowledge in the test battery of Ellis' (2009) study, did not measure implicit knowledge but drew on automatized explicit knowledge. Even though time pressure was imposed and attention was directed to meaning during the EI task, there appeared to be some room for accessing automatized explicit knowledge for advanced L2 learners. Since EI tasks direct learners' attention to meaning, it is certainly a better test of implicit knowledge than form-focused tasks like the timed GJTs. These timed GJTs should be deemed a measure of automatized explicit knowledge because they always require learners to pay attention to forms, which inevitably raises awareness of one's linguistic knowledge (Vafaee et al., 2017). Of course, the level of awareness that each test taker brings to the task may vary depending on his/her background. For instance, native speakers and some L2 learners (e.g., heritage learners) with little experience of learning of a L2 through formal instruction, who presumably possess little explicit knowledge, need to draw on implicit knowledge to perform a GJT, regardless of its being timed or untimed. In contrast, learners with formal instruction may tend to recourse to, or at least attempt to draw on, automatized explicit knowledge. The present study focuses on L2 learners with some formal instruction and hypothesizes that timed GJTs primarily draw on automatized explicit knowledge for them.

THEORETICAL IMPORTANCE FOR DISTINGUISHING BETWEEN IMPLICIT KNOWLEDGE AND AUTOMATIZED EXPLICIT KNOWLEDGE

The distinction between automatized explicit knowledge and implicit knowledge in L2 learners has not been thoroughly researched. Differentiating linguistic knowledge that has no conscious correlate (implicit) from that which involves consciousness (explicit) but has been automatized bears important implications at many levels. From an applied pedagogical perspective, the distinction may be trivial practically (see further discussions in DeKeyser, 2015; Spada, 2015). From a theoretical point of view, however, the distinction is indispensable for tackling two related issues in explicit and implicit learning. The first problem concerns the nature of linguistic knowledge types that L2 learners possess. By postulating automatized explicit knowledge in addition to implicit and (nonautomatized or less automatized) explicit knowledge, it allows for assessing L2 ability through more scrutinized constructs. For instance, it is an empirical question of to what extent L2 proficiency (e.g., measured by standardized tests) can be explained by implicit knowledge, automatized explicit knowledge, and less automatized or nonautomatized explicit knowledge (e.g., Elder & Ellis, 2009).

A more important point is that accurate identification/assessment of distinct types of linguistic knowledge provides essential insight into the second problem, that is, uncovering L2 learning processes. One of the central issues in the SLA field is how explicit/implicit learning leads to the acquisition of implicit knowledge:

the interface issue. Many researchers express different positions as to whether explicit knowledge facilitates the acquisition of implicit knowledge (DeKeyser, 2015; N. C. Ellis, 2005; Ellis, 2008; Hulstijn, 2002; Krashen, 1985; McLaughlin, 1987; Paradis, 2009). The theoretical distinction between automatized explicit knowledge and implicit knowledge, along with valid measurements for them, can advance our understanding of the interface issues in at least two related areas.

Explicit learning processes can be examined in more depth as L2 learning results in a large variability in the degree of automatization in L2 knowledge (e.g., DeKeyser, 1997, 2015). Implicit learning processes can be examined more closely in relation to different types of explicit knowledge. A certain group of L2 learners may first engage in explicit learning and succeed in attaining automaticity; they may utilize automatized explicit knowledge to facilitate the acquisition of implicit knowledge (Suzuki & DeKeyser, 2017). In contrast, a different type of L2 learner possesses explicit knowledge, a large part of which is not automatized at all; these learners may have to deploy implicit learning mechanisms in different ways from the first group. It is also possible that the usefulness of explicit knowledge for implicit learning varies depending on whether explicit knowledge is automatized or not. The clearer operationalization and identification of the constructs are crucial in revealing learning processes of different L2 learner populations through the lens of explicit and implicit learning.

NEW MEASURES OF IMPLICIT KNOWLEDGE: REAL-TIME COMPREHENSION TESTS

Following Suzuki and DeKeyser (2015), the current study proposes that implicit knowledge is drawn upon when test takers *register*[1] specific grammatical structures for real-time comprehension. Examining real-time grammar processing allows us to capture whether learners can deploy their linguistic knowledge with very little lag from the input; they are very unlikely to apply linguistic knowledge consciously (Andringa & Curcic, 2015; Leung & Williams, 2012; Paradis, 2009; Suzuki & DeKeyser, 2015). Only implicit knowledge makes it possible to operate at almost the exact time of occurrence of targeted grammatical structures. More important, measures of implicit knowledge should direct test takers' attention primarily to meaning so that they do not raise awareness about grammatical structures to be targeted. While form-focused tasks are direct measures of grammatical knowledge, implicit knowledge tests are characterized as indirect measures. In what follows, I will introduce three psycholinguistic measures that capture real-time comprehension of grammatical structures, requiring no grammatical judgments on the stimulus sentences. I will first discuss reaction-time measures with focus on a word-monitoring task and a self-paced reading task. After that, I will introduce a still newer method in the L2 field, an eye-tracking while-listening task (i.e., visual-world task).

An increasing interest in a psycholinguistic approach to SLA has developed over the decades, leading to an increasing use of reaction time (RT) to examine online sentence processing in L2 (for a review, see Jiang, 2011). Representative tasks include the word-monitoring task (Granena, 2013; Suzuki & DeKeyser, 2015) and the self-paced reading task (Foote, 2011; Jiang, Novokshanova, Masuda, &

Wang, 2011; Roberts & Liszka, 2013). The advantage of using these RT methods, over form-focused tasks like GJTs, is that we can indirectly measure grammatical sensitivity without asking for grammaticality judgments. In the word-monitoring task, participants listen for a monitoring word and respond to it as soon as they hear it in an auditory sentence by pressing the key on the computer. They pay attention to the meaning of the sentences, rather than to the grammatical forms, because comprehension questions are presented after they hear the sentence. The monitoring word is embedded in an auditory sentence and occurs right after a target grammatical structure. When participants listen for a monitoring word (e.g., to) in an ungrammatical sentence (e.g., The man likes to play basketball), they are expected to slow down to respond to the target word, compared to the one in the grammatical counterpart (e.g., The man likes to play basketball). The RT difference between grammatical and ungrammatical indicates the extent to which participants detect the errors without awareness. The same logic of assessment of online grammatical sensitivity applies to the self-paced reading task where participants read a sentence word by word on the computer (see Instruments).

RT-based research has stood as a gold standard for psycholinguistic studies; however, a more fine-grained measurement technique, a visual-world task, has started to be utilized to capture real-time L2 grammar processing (Grüter, Lew-Williams, & Fernald, 2012; Hopp, 2013; Lew-Williams & Fernald, 2010; Trenkic, Mirkovic, & Altmann, 2014). In the visual-world task, participants see a visual scene consisting of pictures while listening to stimulus sentences with target grammatical structures. By analyzing eye movements during the listening process, it can reveal real-time comprehension of grammatical structures (Sedivy, 2010; Tanenhaus & Trueswell, 2006). The advantages of applying the visual-world paradigm to L2 research are summarized as follows: (a) it requires no ungrammatical sentences (little risk of raising awareness), (b) it is a direct measure of fast and ballistic (unstoppable) linguistic processing in real time, (c) it is simple and can be applied to wider populations, and (d) it enjoys higher ecological validity than RT tasks. All in all, the three tasks, by virtue of measuring real-time comprehension process, should each measure implicit knowledge.

AMOUNT OF L2 EXPOSURE INFLUENCES RELIANCE OF IMPLICIT KNOWLEDGE

In addition to examining the relationships among the language test scores, the current study aims to obtain further evidence for validity of the new implicit knowledge measures. It takes a very long time to acquire implicit knowledge because a large amount of L2 input for specific grammatical forms is required to develop this (Paradis, 2009). Individual differences in the amount of L2 exposure have been found to be related to the acquisition of implicit knowledge (Suzuki & DeKeyser, 2015). The work by Suzuki and DeKeyser (2015) revealed that performance for a measure of implicit knowledge (i.e., the word-monitoring task) was correlated with scores of the implicit learning aptitude (i.e., measured by the serial-reaction time task) *only* among the L2 learners with longer length of residence (LOR) in the immersion context. It is conceivable that L2 learners with more L2 experience are more likely to rely on implicit knowledge stably on the

Table 1. *Task features of the linguistic knowledge measurements*

	Indirect-Implicit Measures			Direct-Explicit Measures		
	Visual-World	Word Monit.	Self-Paced	Timed AGJT	Timed VGJT	Timed SPOT
Data	Fixation Proportion	RT	RT	Accuracy	Accuracy	Accuracy
Real-time processing	Yes	Yes	Yes	Yes/No	Yes/No	No
Focus	Meaning	Meaning[a]	Meaning	Form	Form	Form
Time pressure	No	Yes	Yes	Yes	Yes	Yes
Modality	Aural	Aural	Written	Aural	Written	Written

Note: AGJT, auditory judgment test; VGJT,visual judgment test; SPOT, Simple Performance-Oriented Test; RT, reaction time.
[a]The focus of attention is also directed to the monitoring word.

language tests. Following Suzuki and DeKeyser (2015), the current study recruited Japanese L2 learners who live in Japan and then divided them into two groups based on their LOR. By examining the LOR in the immersion context (a proxy for the amount of L2 exposure), the current study contributes to a better understanding of the measurement and development of implicit knowledge. It can also potentially inform a more stringent participants selection procedure for testing implicit knowledge.

THE CURRENT STUDY

The aim of the current study was to examine the validity of the behavioral measures that could measure implicit knowledge and automatized explicit knowledge separately. Six language tests were developed to assess linguistic knowledge of three Japanese grammatical structures and were administered to 100 Japanese L2 learners. The three indirect, real-time comprehension measures of grammatical knowledge (the visual-world task, the word-monitoring task, and the self-paced reading task) were hypothesized to assess implicit knowledge, whereas the other direct, form-focused measures (the timed auditory GJT, the timed visual GJT, and the timed fill in the blank test called Simple Performance-Oriented Test [SPOT][2]) were hypothesized to draw on automatized explicit knowledge.

As shown in Table 1, the crucial differences between the two types of measures lie in (a) real-time sentence processing and (b) focus of attention. All three online measurements assess whether test takers can incrementally process the sentence while their attention is focused on the meaning of the sentences. They are less likely to use linguistic knowledge consciously because their real-time grammatical processing is examined within a time window of few hundred milliseconds (Andringa & Curcic, 2015; Leung & Williams, 2012; Paradis, 2009; Suzuki & DeKeyser, 2015). In contrast, the two types of GJTs and the SPOT require them to focus on form or grammatical target points under time pressure. Even if the time pressure is imposed on them, they are more likely to use explicit knowledge because the tasks inherently predispose them to focus on form (Vafaee et al., 2017).

There are some differences between the GJTs and the SPOT. First, the amount of attention to form may be greater in the SPOT than in the GJTs. In the SPOT (i.e., fill in the blank test), learners have to focus on specific grammatical structures to fill in the blanks, whereas in the GJTs they do not know whether and where grammatical errors are embedded in each test item and need to search for grammatical errors. Second, the SPOT might not have imposed as strong incentives to respond quickly as the GJTs to complete the task, because a longer time limit was set on each test item in the SPOT than in the GJTs (see Methods). It is still possible that some learners make grammatical judgments in real time on the timed GJTs; the requirement for "real-time processing" is less certain for the GJTs (see Table 1).

In order to validate the measurements for implicit knowledge and automatized explicit knowledge, CFA was conducted to assess construct validity. In contrast to exploratory factor analysis, CFA is a better approach to estimate relationships among measured variables because it allows for identifying latent constructs by taking into account the measurement errors. The CFA procedures consist of (a) initial model specification, (b) model evaluation, and (c) rival model comparison. In the initial model specification, a CFA model is specified in advance based on prior theories. Here, the two-factor model was hypothesized (see the left panel in Figure 1).

In model evaluation, the model is then tested with the gathered data and evaluated by a goodness of fit. The identified model is then assessed for parameter estimates such as factor loadings, error variances, and correlations between factors. Each parameter provides important information to examine validity. Factor loadings represent the amount of variance in a measured variable (e.g., timed auditory GJT) explained by the factor. For instance, high factor loadings of timed auditory GJT, timed visual GJT, and timed SPOT suggest that these three measures measure a common theoretical construct (e.g., automatized explicit knowledge). In other words, they serve as supporting evidence for *convergent validity* or the extent to which measured variables are related (Campbell & Fiske, 1959). In contrast, *discriminant validity* refers to the extent to which a latent factor (e.g., implicit knowledge) discriminates from other latent factors (e.g., automatized explicit knowledge). Discriminant validity can be evaluated by examining the relation between the factors. A weak relationship between the two factors that were hypothesized in the current study indicates the dissociation between implicit and automatized explicit knowledge.

Further evidence for the discriminant validity can be evaluated by a rival model comparison. As shown in the right panel in Figure 1, the one-factor model can be specified as a rival model against the two-factor model.[3] The one-factor model can be plausible because all the language tests assess a single type of linguistic knowledge. If the two-factor model is found to be better than the one-factor model, it suggests that all six measures are not tapping into a single construct but two distinct constructs, hence supporting the discriminant validity.

The current study takes a further step to evaluate the construct validity by conducting multitrait-multimethod (MTMM) analysis (Widaman, 1985). The key advantage of MTMM analysis is that it assesses the extent to which the traits (i.e., latent constructs) were measured validly by taking into account the method effects

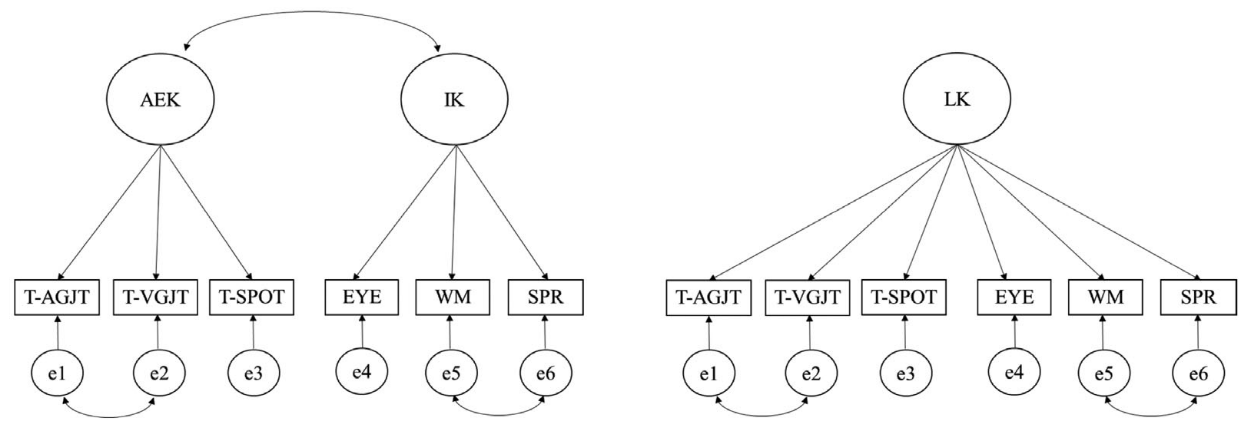


Figure 1. Confirmatory factor analysis models: (left) two-factor model and (right) one-factor model. AEK, automatized explicit knowledge; IK, implicit knowledge; LK, linguistic knowledge; T-AGJT, timed auditory grammaticality judgment test; T-VGJT, timed visual grammaticality judgment text; T-SPOT, timed SPOT; EYE, visual-world task; SPR, self-paced reading task; WM, word-monitoring task.

(see, e.g., Bachman & Palmer, 1982, for applications of MTMM in L2 learner populations). The current study utilized a pair of measurements that shared very similar methods (e.g., the visual GJT and the auditory GJT). This leads to a potential threat to the validity because some portion of correlations between similar measures can be simply due to the similarity in methods (method effects) not to the underlying common construct. MTMM analysis allows for assessing the extent to which variance in the measurements could be attributed to traits versus to methods (Podsakoff, MacKenzie, & Podsakoff, 2012). Specifically for the purpose of this study, the present study addressed the construct validity as to whether the traits (automatized explicit and implicit knowledge) could be measured rather than the method effects (RT and GJT measures). Error covariance was imposed on two pairs of measurements that shared similar methods (see Figure 1): the word-monitoring task and the self-paced reading task, which utilized similar RT measures while listening or reading for comprehension, and the visual GJT and the auditory GJT, which shared the same procedure except for the modality difference.

RESEARCH QUESTION AND HYPOTHESES

The current study addressed the question whether the test battery measures the constructs of implicit knowledge and automatized explicit knowledge separately. The two-factor model was evaluated by five criteria for construct validity. First, the two-factor model was examined as to whether it fits the current data set. Second, it was investigated to what extent the measurements assessed either automatized explicit knowledge or implicit knowledge construct (convergent validity). Third, it was investigated the extent to which the set of measurements for implicit knowledge and those for automatized explicit knowledge were dissociated (discriminant validity). This discriminant validity for the two factors was tested by (a) computing correlations between the two factors and (b) comparing the two-factor model and the one-factor model. Fourth, a MTMM analysis was performed to assess traits and method effects. Fifth, the study examined if the amount of L2 exposure in the immersion setting, estimated by the LOR, moderated the results of the four criteria above. For these criteria, five hypotheses were put forth:

1. *Hypothesis 1:* The data structure of the six measurements demonstrates a good fit to the two-factor model.
2. *Hypothesis 2:* The factor loadings are strong and significant (systematic) for automatized explicit knowledge and implicit knowledge (convergent validity).
3. *Hypothesis 3a:* The relationship between the two latent factors is insubstantial (discriminant validity).
   Hypothesis 3b: The data structure of the six measurements demonstrates a poor fit to the one-factor model (discriminant validity).
4. *Hypothesis 4:* The error covariance between the similar measurement methods is nonsignificant or smaller than the covariance between the measurements for the traits (method effects).
5. *Hypothesis 5:* The results from L2 learners who received long-term exposure in the immersion setting confirm Hypotheses 1–4 more convincingly than the results from L2 learners with less exposure.

Table 2. *Background Information of the second language learners*

	Age at Testing	Age of Arrival	Onset of Instruction	LOR (months)	Length of Instruction (months)
Whole group (n = 100)					
Mean	25.97	21.36	19.01	47.29	41.11
SD	4.47	2.66	2.25	27.71	17.44
Range	19–47	17–30	13–27	24–197	3–84
Short-LOR group (n = 48)					
Mean	23.88	21.21	18.69	30.13	41.54
SD	2.72	2.63	1.82	4.33	17.16
Range	19–32	17–29	13–24	24–38	6–72
Long-LOR group (n = 52)					
Mean	27.90	21.50	19.31	63.13	40.71
SD	4.91	2.72	2.57	30.66	17.84
Range	22–47	17–30	13–27	39–197	3–84

Note: LOR, length of residence.

METHODS

Participants

One hundred Japanese L2 learners (29 male, 71 female), whose first language was Chinese, were recruited in Tokyo and the surrounding Kanto area. Four requirements had to be met by L2 learners in order to participate in the study: proficiency, age of arrival in Japan, LOR, and educational background. First, only advanced-level Japanese L2 learners were recruited. They were screened for Japanese proficiency, which had to be equivalent to or higher than N1 in the standardized Japanese Language Proficiency Test.[4] Second, I only focused on late L2 learners, who arrived in Japan after the age of 17. Third, their LOR in Japan was 2 years or longer. This cutoff point for LOR was roughly based on the previous findings that implicit knowledge seems to be exhibited most efficiently in online measurements (i.e., the word-monitoring task) when L2 learners have been immersed in the target country for 2.5 years of residence or longer (Suzuki & DeKeyser, 2015). Fourth, participants possessed at least a bachelor's degree or were enrolled in a 4-year college at the time of testing. The sampled population consisted of undergraduate students (n = 34), MA students (n = 40), PhD students (n = 14), and office workers (n = 12) at the time of testing. Forty-three out of 100 participants majored in Japanese language studies (i.e., Japanese or Japanese education as a foreign/second language) in undergraduate studies; 27 out 61 participants with an MA degree or currently seeking one in Japanese language studies; and 5 out of 14 participants with a doctorate degree or who are pursuing one in Japanese language studies. The rest of the participants' major varied (e.g., economic, architecture, engineering, management, law, psychology, physics, and liberal arts). Background information about the L2 learners is presented in Table 2.

The whole group was split in half by using the median LOR of 39 months (see Table 2). According to independent *t* tests, the two groups were significantly different in terms of LOR and age at testing ($p < .001$). The other factors (age of arrival, onset of instruction, and length of instruction) were not different ($p > .05$). Fifty-one native speakers (NSs) were also recruited to serve as a baseline for the linguistic knowledge tasks (see the Analysis section).

Target structures

Three Japanese linguistic structures were tested across the six language tests: transitive/intransitive verb pairs, classifiers, and locative particles (*ni/de*). These structures were chosen because they generate some prediction of upcoming information, which can be demonstrated by the visual-world task. All target structures are usually explicitly taught in beginner-level Japanese classes.

Transitive–intransitive verb pairs. Sixteen transitive–intransitive verb pairs were chosen (Jacobsen, 1992). The pairs share the stem, but morphological markings distinguish transitive from intransitive. For instance, the transitive verb *war-u* (to break) has the intransitive counterpart *war-eru* (to be broken). A theme is discernible by the object-marking particle *o* for the transitive verb (e.g., *sara-o waru*; someone breaks the dish). For the intransitive verb, the theme should be marked with the subject-marking particle *ga* rather than *o* (e.g., *sara-ga wareru*; the dish got broken). Note that action doer is implied in the transitive verb.

Classifiers. Eight classifiers were chosen and matched with 4 nouns; there were 32 classifier–noun pairs. For instance, *chaku* is a counter for clothes as in *go-chaku no doresu (five-CHAKU-Genitive dress; five dresses).* Although some classifiers are shared between Japanese and Chinese, we chose the classifier–noun pairs that were not shared in order to avoid mere transfer from Chinese to Japanese (see online-only supplementary material Appendix A).

Locative particles: Ni/De*.* The particles *ni* and *de* are multifunctional case markers, and the usage for locations was focused on in the current study. In particular, *de* indicates the place where an action takes place (e.g., *toshokan-de benkyousuru*; study in the library), whereas *ni* is used to indicate the place where a thing or a person exists (e.g., *toshokan-ni iru*; I will be in the library). It has been found that Chinese speakers tend to overuse *ni* for *de* (Hasuike, 2004). Not all of the usage for *ni* is difficult, and a relatively easier usage is expressing destination with motion verbs (e.g., *cafe ni hairu*; enter the cafe). In sum, action verbs agree with the location particle *de*, static verbs with the location particle *ni*, and motion verbs with the destination particle *ni*.

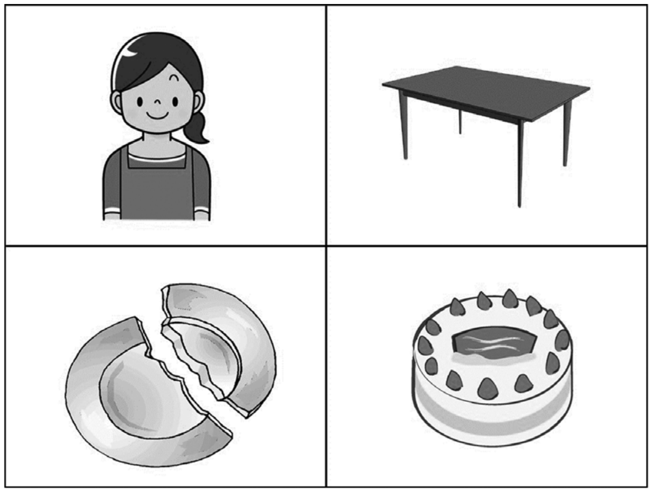
Instruments

Visual-world paradigm. In the visual-world task, participants were first presented with a scene consisting of four pictures on the computer screen for 5.5 s. They then listened to sentences while their eye movements were being tracked, using an EyeLink-II system (SR Research, Osgoode, Ontario, Canada) with a sampling rate of 500 Hz. Participants were presented with a total of 64 trials: 48 critical trials and 16 filler trials. Sixteen trials were prepared for each of the three linguistic structures tested, and participants heard two sentences for each trial. The critical sentence was always presented as the first sentence so that participants were not influenced by any information from the previous sentence (16 trials $\times$ 3 structures $=$ 48 sentences). The second sentence acted as a filler to divert the participants' attention from the critical sentence and to avoid revealing the purpose of the study (48 filler grammatical sentences). There were also 16 filler trials, resulting in 32 filler grammatical sentences. All trials were presented in semirandomized order such that the same trial type never occurred more than twice in a row. The location of the four objects on display was rotated across trials. After each trial, a yes/no comprehension question was asked to ensure that participants' attention was focused on the meaning of the sentence (cf. Dussias et al., 2013). [Q1] Half of the questions asked about the critical sentence, and the other half about the filler sentence. Two practice trials were given to familiarize the participant with the procedure of the task.

The display always involved a target object and a competitor object. There were two types of trials for each target structure: target trials (where the target object was mentioned in the critical sentence) and the competitor trials (where the competitor object was mentioned). As shown in Figure 2, each display for transitive/intransitive structures consisted of a person (e.g., the mother), a contrast object (e.g., the table), a theme (e.g., the broken dish), and a distractor. The person was defined as a target, whereas the contrast object was defined as a competitor. Two types of critical trials were created: transitive and intransitive trials.

The first part of both sentences always followed the same form: *NP1-ACC-transitive verb-iru-no-wa-adverb-NP2* (It is NP2 that TRANSITIVE-VERB NP1) and *NP1-SUB-intransitive verb-iru-no-wa-adverb-NP2* (The reason is NP2 why NP1 INTRANSITIVE-VERB), where NP1, ACC, and NP2 are noun phrase 1, accusative, and noun phrase 2, respectively. NP2 was always a person (e.g., the mother) in the transitive trials (defined as target trials), whereas it was always a contrast object (e.g., the table) in the intransitive trials (defined as competitor trials). The eye movements were analyzed from the onset of the case marker (*ga* or *o*). If participants were sensitive to the transitivity of the verb, then looks to the target (e.g., mother) would be greater in the target trials than in the competitor trials. This is because a segment of NP-ACC and *te*-form of a transitive verb (i.e., *osara-wo watte*) implied an action doer. The task design for the other two structures is described in online-only supplementary material Appendix B.

We were primarily interested in looks to the two possible locations, coded as targets and competitors. Before the primary analyses, each time window was shifted 200 ms after the linguistic cues in the speech stream to account for the time it takes to generate a saccadic eye movement (Matin, Shao, & Boff, 1993). In order to compute a "sensitivity index" for individuals, "target advantage (TA) scores" were first computed separately for target trials and competitor trials as follows: target looks divided by the sum of target looks and competitor looks. TA scores were then standardized (z transformed) across the three structures, and the sensitivity index was computed by the TA difference scores as follows: TA in the target trials – TA in the competitor trials. The higher sensitivity score indicated more developed linguistic knowledge. These sensitivity scores were computed after time locking to 200 ms



(1) Transitive trials (Target trials):

Osara wo watte iru no wa soko ni iru okaasan desu.
Dish-ACC breaking be NOMINALIZER TOP there-LOC exist mother be.
(It is the mother that is breaking the dish.)

(2) Intransitive trials (Competitor trials):

Osara ga warete iru no wa soko ni aru teeburu kara ochite shimatta kara desu.
Dish-SUB breaking be NOM. TOP there-LOC exist table from fall off because.
(The dish is broken because it fell off the table.)

Figure 2. Visual scene and critical sentences for transitive/intransitive structure.

from the data-drive onset (see online-only supplementary material Appendix C for details).

Word-monitoring task. In the word-monitoring task, participants were instructed to listen to a sentence for a target word and to press a button as soon as they identified it in the spoken sentence. The target word remained on the screen until the response was made. A yes/no comprehension question appeared on the screen, so that participants' attention was directed to the sentence meaning as well as the target word. For instance, a sample sentence targeting transitive structure is presented below.

[Target Word: *mazeru*]
*Ao to kiiro no enogu o/*ga mazeru to, kirei na mimdori ni naru.*
Blue and yellow paint-ACC/SUB mix if, beautiful green become
When you mix blue and yellow paints, it becomes beautiful green.

The target sentence always included a segment of the case-marking particle (*ga* or *o*) and a verb (transitive or intransitive). The target word was always the verb following the particles (*ga* or *o*). The "sensitivity index" was computed by the RT difference scores (ungrammatical RT – grammatical RT) across the three structures, after the average RTs of grammatical and ungrammatical items were standardized (z scores) in order to treat the sensitivity across the target structures equally. The magnitude of this sensitivity index is used to index how developed one's implicit knowledge is (Granena, 2013; Suzuki & DeKeyser, 2015).

The list of stimulus sentences included 48 target sentences (16 for each structure, half grammatical and half ungrammatical) and 48 grammatical filler sentences. Half of the items for each condition were followed by a yes/no comprehension question. The ratio was kept equal between a positive response and a negative response. Sample stimulus sentences for the other structures are presented in online-only supplementary material Appendix D.

Self-paced reading task. In the self-paced reading task, participants were asked to read a sentence word by word as quickly as possible while paying attention to its meaning to answer a comprehension question accurately. The first word of a sentence appeared on the left side of the screen, and when the button was pressed, the next word appeared to the right of the preceding word, which disappeared upon the presentation of the following word (moving-window presentation). When participants read the final word followed by the period, they pressed a second key to continue to either the next test item or a comprehension question. Words were presented in Japanese characters in chunks consisting of a clause or *bunsetsu* (i.e., content word + function word). For example, a sample sentence with the transitive structure is presented below (a slash indicates a unit of presentation).

[Region 1 = *mazeru to*, Region 2 = *ii*]

Uta no gurupu o/	*tsukuru*	*tokini/*	*danshi*	*to/*	*joshi o(*ga)/*
Singing group-OBJ/	make	when	boy	and	girls-OBJ

mazeru to/	*ii/*	*baransu*	*ni/*	*naru to omou/*
mix	if	good	balance	becomes think

When you form a singing group, I think it makes a good balance if you mix boys and girls.

The region of interest where RTs were compared between grammatical and ungrammatical sentences was at the critical word where the error occurred in the ungrammatical sentences (Region 1). This word was located in the same position as that in the word-monitoring task so that the effects could be compared fairly between the word-monitoring task and the self-paced reading task. RTs of the word immediately following the critical word (Region 2) were also included to capture spillover effects (Mitchell, 1984). In a similar way to the word-monitoring task, the sensitivity index was computed for individuals as z-standardized RT scores (ungrammatical RT – grammatical RT) at Regions 1 and 2 combined.

As in the word-monitoring task, a list of stimulus sentences included 48 target sentences (16 for each structure, half grammatical and half ungrammatical) and 48 grammatical filler sentences. Half of the items for each condition were followed by a yes/no comprehension question. The ratio was kept equal between a positive response and a negative response. Sample stimulus sentences with the other structures are presented in the online-only supplementary material Appendix E.

Timed auditory GJT. In the computer-delivered timed auditory GJT, participants listened to an aural stimulus sentence and indicated whether each sentence was grammatical or ungrammatical by pressing a response button. They were asked to press a key as soon as an error was detected in the sentence. The time limit imposed on the task was 10 s for each item. Responses that were longer than certain time limits were then dealt with after administering the test (see Data Analysis section for details). The stimulus sentences consisted of 48 target sentences (16 for each structure, half grammatical and half ungrammatical) as well as 16 grammatical filler sentences. Before the actual test, participants took a practice session. The percentage accuracy score was calculated for all the items. One item in the auditory GJT was excluded from the analyses because the sentence was not unambiguously grammatical or ungrammatical (58% in NS accuracy rate).

Timed written GJT. As in the timed auditory GJT, the timed visual GJT was also administered on a computer. The procedure was identical to the one in the timed auditory GJT except for the modality of the stimulus sentences. Participants were presented with a written sentence on a screen and asked to indicate whether each sentence was grammatical or ungrammatical by pressing a response button as quickly as possible. They were allowed to press the key while the sentence was played when the error was detected within the sentence. The time limit imposed on the task was 10 s for each item. The stimulus sentences consisted of 48 target sentences (16 for each structure, half grammatical and half ungrammatical) as well as 16 grammatical filler sentences. The percentage accuracy score was calculated for all the items. One item in the visual GJT was excluded from the analyses because the sentence was not unambiguously grammatical or ungrammatical (68% in NS accuracy rate).

Timed SPOT (fill in the blank test). In the timed SPOT, the participants were presented with a single sentence with some blanks on the computer screen. Then, they had to fill in the blank with Japanese characters on the answer sheet as quickly as possible. A blank was left in each sentence to specifically target one of the linguistic structures. Once they filled in the answer on the sheet, they pressed a computer button to move on to the next item. Participants were told to respond as quickly as possible. The time limit for each test item was accidentally set to 100 s, instead of 10 s (see Data Analysis section). The number of characters to be filled in the sentence was indicated by the number of blank circles in the sentence (see sample items in online-only supplementary material Appendix F). A syllabic *hiragana* character was used to fill in the blanks. The stimulus set consisted of 48 target sentences (16 for each structure) and 16 filler sentences. The percentage accuracy score was calculated over all items for the target sentences.

Procedure

Participants were tested individually in a soundproof booth. After the consent form and the background questionnaire, the linguistic tasks were administered in fixed order from the most implicit linguistic tasks to the more explicit: the visual-world task, the word-monitoring task, the self-paced reading task, the timed auditory GJT, the timed visual GJT, and the timed SPOT. Before taking each task, participants were presented with several practice items to familiarize them with the procedure. All the stimulus sentences were different across the tasks in order to reduce practice effects. They were presented in a fixed semirandom order in each task, interspersing different types of stimulus sentences, in order to conceal the purpose of the study. It took approximately 2 hr to complete the tasks, and participants were given two 3-min breaks, one after the visual-world task and another after the self-paced reading task.

Data analysis

Real-time comprehension tasks. For all three implicit knowledge tests (real-time comprehension tasks), data cleaning procedures were conducted. Specifically, accuracy of the comprehension questions was computed. A participant whose error rate was higher than 25% would be excluded from further analysis to ensure that each individual was paying attention to meaning (Jiang et al., 2011). None of the participants scored below 75%; all participants' eye-movement and RT data were analyzed. More detailed results from data cleaning procedures are presented in online-only supplementary material Appendix G.

Timed form-focused tasks. Previous studies like R. Ellis (2005) and Bowles (2011) set the time limit for *presenting* each sentence based on the NSs' average response time plus an additional 20% of the time for each sentence. A more lenient time pressure was imposed on the current tasks: 10 s across all the test items. Instead of imposing a strict time-out for duration of sentence presentation, L2 learners' responses were screened after the data was collected. If the response time was not within a certain time limit based on the NSs' RTs, those responses were scored as incorrect. Initial review of data revealed that around 15%–30% of the responses would be discarded *even for NSs' responses* in the three form-focused tasks when we imposed the 20% + NSs' RT for each item. It seemed more reasonable to impose time pressure in which most NSs can perform the task accurately. We decided to identify a different percentage value so that 90% of the NSs' responses were scored correct. In other words, percentages to be added to the NSs' mean RT were determined such that the NSs' mean error rate of the total score was kept less than 10%. The cutoff percentages that retained 90% of NSs data were mean RTs + 50% for the auditory GJT, mean RTs + 120% for the visual GJT, and mean RTs + 50% for the SPOT. These cutoff points were used to score the responses of L2 learners in the three tests.

Data summary: Missing data and data transformation

Before presenting the results for L2 learners, native speakers' performance on the six language tests was checked (see online-only supplementary material Appendix H). They showed sensitivity to the manipulation of stimulus sentences in the meaning-focused tests (visual-world, word-monitoring, and self-paced reading

Table 3. *Descriptive statistics for the language tests by second language learners*

	N	Possible Max	*M*	*SD*	Min	Max	95% CI	Cronbach α
Eye[a]	100	—	0.01	0.09	– 0.26	0.24	[-0.01, 0.03]	—
WM[a]	100	—	22	54	– 111	162	[11, 33]	0.91
SPR[a]	100	—	36	90	– 198	351	[18, 54]	0.96
T-AGJT[b]	100	100	43.43	12.12	14.58	76.19	[0.41, 0.46]	0.67
T-VJGT[b]	100	100	30.64	16.28	2.08	82.74	[0.27, 0.34]	0.85
T-SPOT[b]	99	100	27.13	23.37	0	91.67	[0.20, 0.29]	0.95

Note: Eye, visual-world task; WM, word-monitoring task; SPR, self-paced reading task; T-AGJT, timed auditory judgment test; T-VGJT, timed visual judgment test; T-SPOT, timed Simple Performance-Oriented Test.
[a]The values for the online comprehension tasks indicate sensitivity index.
[b]The values for the form-focused tasks indicate percentage accuracy score.

tasks) and high accuracy (all above 90% in accuracy) in the form-focused tasks (auditory and written GJTs and SPOT).

Descriptive statistics for all the measures performed by L2 learners are presented in Table 3. L2 learners showed no sensitivity to the target structures in the visual-world task, whereas they demonstrated some sensitivity in the word-monitoring and the self-paced reading tasks (see online-only supplementary material Appendix I for details). Their performance on the form-focused tasks was low; they scored highest on the timed auditory GJT, followed by the timed visual GJT, and then the timed SPOT.

Reliability indices were all above .65 and deemed acceptable (Loewenthal, 2004). The timed auditory GJT showed lower reliability (.67) than the other form-focused tasks in the test battery perhaps because the test takers had only one chance to listen to a spoken stimulus sentence. The internal consistency (e.g., Cronbach α) was not computed for the visual-world task in the current study because no standard procedure exists for estimating internal consistency of the visual-world task; one promising approach is to examine test–retest reliability (Farris-Trimble & McMurray, 2013). Since the test–retest reliability was not available in the current study, the current findings should be interpreted with caution.

Prior to the CFA and MTMM analyses, tests of univariate normality were examined for the six test scores. The total scores of the T-SPOT were positively skewed; square root transformation was applied to reduce skewness. Based on the standardized coefficients of skewness and kurtosis (z scores), all the variables met the assumption of univariate normality ($p > .05$). Multivariate normality of the score distribution was examined by Mardia's coefficient. The coefficients (chi-square) were 1.648 ($p = .439$) for all the six tests and 0.007 ($p = .996$) for the five tests, both of which met the assumption of multivariate normality. Out of 100 participants, only 1 participant had missing cases in T-SPOT. Since this person was the only one who had a missing case in the language tests, this person was excluded from the analyses.

CFA and MTMM analysis

The two hypothesized CFA models (one-factor and two-factor models) were entered into the CFA analyses (Figure 1). All the analyses were implemented in the software package LISREL 9.1 (Jöreskog & Sörbom, 2013). Five hypotheses were evaluated. The models were evaluated statistically with a maximum likelihood method to estimate the model parameters (Hypothesis 1). Multiple fit indices were jointly used to assess the model fit in addition to the chi-square statistics (Brown, 2006; Hoyle & Panter, 1995). The following three categories of fit indices were utilized to assess the overall goodness of fit of the CFA models: absolute fit indices (standardized root mean square), incremental fit indices (the comparative fit index and the Benter–Bonnet nonnormed fit index), and fit indices adjusting for model parsimony (root mean square error of approximation). According to the findings of simulation studies conducted by Hu and Bentler (1999), a good fit between the target model and the observed data (maximum likelihood estimation) was obtained in instances where standardized root mean square residual values were below 0.09, root mean square error of approximation values were below 0.06, and comparative fit index and Benter–Bonnet nonnormed fit index were above 0.96. Based on these empirically derived criteria, each of the models was assessed to exhibit one of three levels of fit: good fit, marginal fit, and poor fit. When the indices in two or three out of three categories met the criteria above, the model was considered to be a good fit (Hu & Bentler, 1999). When none of the fit indices reach the criteria, the model was considered to be a poor fit.

In order to seek evidence for convergent validity (Hypothesis 2), the magnitudes and significance of the factor loadings were examined. The discriminant validity was assessed by the correlation between the two latent factors (Hypothesis 3a). In addition, the discriminant validity was also evaluated by comparing the one-factor and two-factor models by the goodness of fit testing indexed by the chi-square statistics as the two models were nested (Hypothesis 3b). A correlated uniqueness model, which is an alternative MTMM approach (Brown, 2006), was constructed to determine the extent to which variance in the measurements could be attributed to latent constructs of linguistic knowledge (traits) and to specific methods (Hypothesis 4). This model correlated the error between the timed visual GJT and the timed auditory GJT and the one between the word-monitoring task and the self-paced reading task.[5] Since the model involved the two traits and two methods, the factor loadings on the same trait factor were constrained to equality (Brown, 2006, p. 220). Finally, the same analyses above were conducted separately for the short-LOR and the long-LOR groups (Hypothesis 5).

RESULTS

Pearson's correlation coefficients will be presented first among the six language test scores, followed by the results from the two competing CFA models with the whole group, short-LOR group, and the long-LOR group. After that, results from MTMM analyses will be presented.

Table 4. *Intercorrelations of the language tests (whole group, n = 99)*

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

Note: Eye, visual-world task; WM, word-monitoring task; SPR, self-paced reading task; T-AGJT, timed auditory judgment test; T-VGJT, timed visual judgment test; T-SPOT, timed Simple Performance-Oriented Test.
$*p < .05$. $**p < .01$.

CFA

Table 4 shows the correlation matrix for the six linguistic test scores for the whole group of L2 learners. Significantly moderate relationships were found among the timed form-focused tasks ($.508 < r < .681$), whereas the correlations among the three online tests were weak, and the only significant relationship among the online measures, between the word-monitoring and the self-paced reading tasks, was weak ($r = .261$, $p = .009$). Unexpectedly, the visual world task was significantly correlated only with T-SPOT, possibly because both tests did not use any ungrammatical sentences.

Both two-factor and one-factor models produced a good fit (see Table 5). A chi-square difference test was conducted to compare the two-factor and the one-factor models. The two-factor model fit better than the one-factor model at the descriptive level, but the difference was not significant, $\chi^2_{\text{difference}} = 0.897$, $df = 1$, $p = .344$. Figure 3 presents both models with factor loadings and significant correlated errors. In the two-factor model, the two latent factors were moderately correlated ($r = .47$, $p = .069$). Factor loadings for automatized explicit knowledge were high and significant, whereas those for implicit knowledge were much lower and the path to the visual-world task (EYE) was only marginally significant. For the one-factor model, the factor loadings for automatized explicit knowledge were identical to the two-factor model, but all the factor loadings for implicit knowledge were lower than the two-factor model. This partially supported that the two-factor model was more plausible than the one-factor model, and the latent factor largely contributes to the form-focused tasks.

In order to investigate how the amount of L2 experience changes the validity of the tests, CFAs were conducted separately for the two subsets. The correlation matrix is presented for the short-LOR and long-LOR groups in Table 6. The form-focused tasks converged to a similar extent for the whole group both in the short-LOR group ($.515 < r < .691$) and in the long-LOR groups ($.534 < r < .626$). While there were no meaningful relationships among the three online tasks in the short-LOR group ($–.129 < r < .100$), the online measures were correlated more highly with each other in the long-LOR group than in the whole group ($.237 < r < .343$).

Table 5. *Fit indices for confirmatory factor analysis models (two-factor and one-factor models) and MTMM models*

	Model	*df*	χ^2	*p*	NNFI	CFI	SRMR	RMSEA [90% CI]	Fit
Whole ($n = 99$)	Two factor	7	6.043	.535	1.019	1	0.036	0 [0–0.113]	Good
	One factor	8	6.940	.543	1.018	1	0.044	0 [0–0.107]	Good
	MTMM	9	9.060	.432	0.999	0.999	0.070	0.008 [0–0.114]	Good
Short LOR($n = 47$)	Two factor				Improper solution				Poor
	One factor	9	4.894	.844	1.159	1	0.055	0 [0–0.094]	Good
	MTMM				Improper solution				Poor
Long LOR($n = 52$)	Two factor	8	7.527	.481	1.015	1	0.071	0 [0–0.156]	Good
	One factor	9	17.328	.044	0.758	0.855	0.116	0.133 [0.022–0.227]	Poor
	MTMM	9	10.622	.303	0.953	0.972	0.097	0.059 [0–0.173]	Good

Note: MTMM, multitrait-multimethod; NNFI, Benter–Bonnet nonnormed fit index; CFI, comparative fit index; SRMR, standardized root mean square residual; RMSEA, root mean square error of approximation; LOR, length of residence. The cutoff values for good fit: SRMR < 0.09, RMSEA < 0.06, and CFI and NNFI > 0.96.

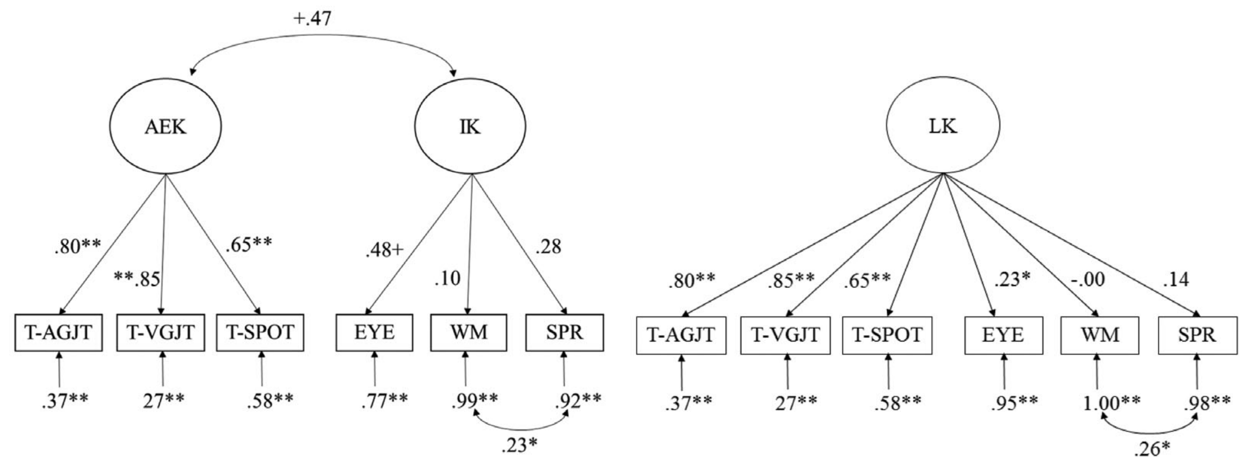


Figure 3. (Left) Two-factor model and (right) one-factor model in the whole second language group ($n = 99$). Standardized coefficients: $+p <$ .10, $*p < .05$, $**p < .01$.

Table 6. *Intercorrelations of the language tests for the short-LOR group (n = 47) and long-LOR group (n = 52)*

	Eye	WM	SPR	T-AGJT	T-VGJT	T-SPOT
Short-LOR group						
Eye	—	−.129	− .057	.146	.217	.170
WM		—	.100	.130	− .010	.165
SPR			—	.142	.137	.128
T-AGJT				—	.691**	.515**
T-VGJT					—	.539**
T-SPOT						—
Long-LOR group						
Eye	—	.237	.343*	.158	.131	.266
WM		—	.270	.010	− .157	.018
SPR			—	.173	.012	.077
T-AGJT				—	.626**	.534**
T-VGJT					—	.567**
T-SPOT						—

Note: LOR, length of residence; Eye, visual-world task; WM, word-monitoring task; SPR, self-paced reading task; T-AGJT, timed auditory judgment test; T-VGJT, timed visual judgment test; T-SPOT, timed Simple Performance-Oriented Test.
$^*p < .05$. $^{**}p < .01$.

The two CFA models were statistically evaluated. For the short-LOR group, the two-factor model failed to converge, and the one-factor model fit the data set well with acceptable fit indices (see Table 5). For the long-LOR group, in contrast, the two-factor model fit the data significantly better than the one-factor model, $\chi^2_{\text{difference}} = 9.801$, $df = 1$, $p = .002$. While the one-factor model yielded a poor fit in all the indices, the two-factor model indicated a good fit (see Table 5). The factor loadings for the good-fit models are presented for the short-LOR (one-factor model) and long-LOR group (two-factor model) in Figure 4.

For the one-factor model of short-LOR group, factor loadings from the measurements hypothesized to assess automatized explicit knowledge were consistently high, but the loadings from the measurements hypothesized to assess implicit knowledge were as low as the whole-group results, suggesting that the L2 learners relied on automatized explicit knowledge more. For the two-factor model of the long-LOR group, factor loadings for the implicit knowledge factor were higher than in the model for the whole group, in addition to the high factor loadings for the automatized explicit knowledge. The covariance between automatized explicit knowledge and implicit knowledge was lower in the long-LOR group ($r = .22$, $p = .258$), suggesting that the two latent factors were more distinct in the long-LOR group than in the whole group.

MTMM analysis

The fit indices of the correlated uniqueness model indicated a good fit for the whole group of L2 learners (see Table 5). As shown in Figure 5, the model results showed that all the trait (factor) loadings were statistically significant ($p < .05$). As in the

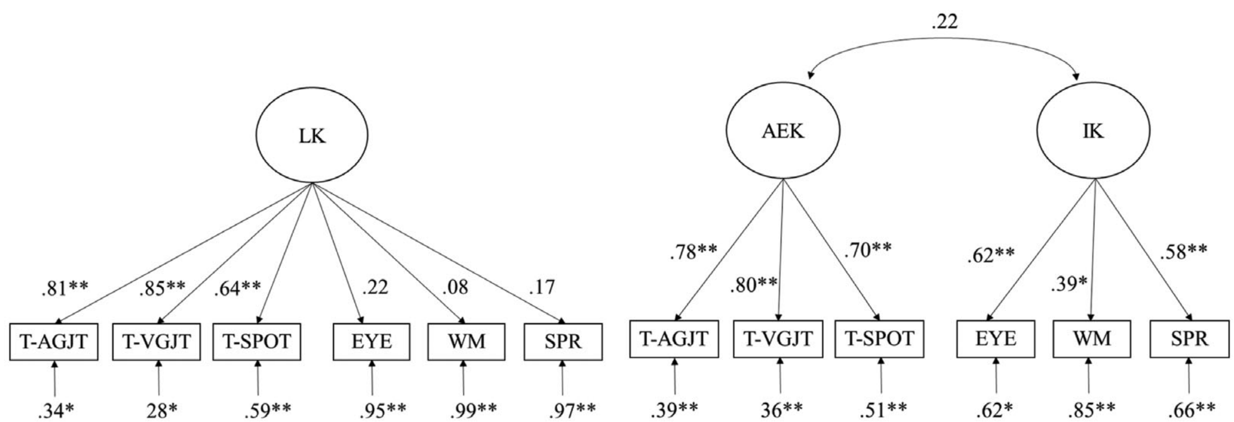


Figure 4. (Left) One-factor model for short-length of residence group ($n = 47$) and (right) two-factor model for long-length of residence group ($n = 52$). Standardized coefficients: $^{*}p < .05$, $^{**}p < .01$.

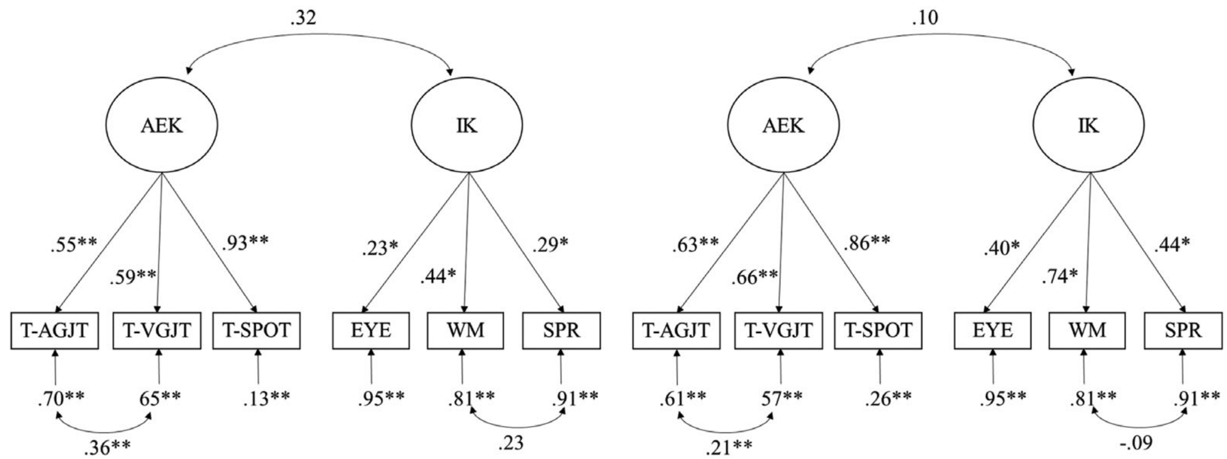


Figure 5. (Left) Multitrait-multimethod models for whole group ($n = 99$) and (right) two-factor model for long-length of residence group ($n = 52$). Standardized coefficients: $^{*}p < .05$, $^{**}p < .01$.

CFA models, the factor loadings were moderate to large in the automatized explicit knowledge measures (range = .55–.93), whereas the trait loadings for implicit knowledge were small to moderate (range = .23–.44). A small and nonsignificant correlation between the two traits was found (r = .32, p = .175). The presence of method effects was examined by the correlated uniqueness (errors) among the similar methods. Although the correlated uniqueness was significant between the visual GJT and the auditory GJT (r = .36, $p < .001$), its magnitude was smaller than any of the trait (factor) loadings from the two GJTs (.55 and .59). The correlated uniqueness between the word-monitoring task and the self-paced reading task was not significant (r = .23, p = .364), and its magnitude was also smaller than either of the trait loadings (.44 and .29). Method effects evaluated in the MTMM model are marginal, indicating that the set of measurements estimated traits reliably with little influence from the method effects.

The same analysis was conducted on the short-LOR group and the long-LOR group, respectively. The model resulted in an improper solution for the short-LOR group; the model for the long-LOR group indicated a good fit of the model with two of the three types of acceptable fit indices (see Table 5).[6] As shown in Figure 5, the model results showed that all the trait factor loadings were statistically significant ($p < .01$). The magnitude of the trait loadings was medium to large, both for the automatized explicit knowledge measures (range = .63–.86) and for the implicit knowledge (range = .40–.74). A nonsignificant negligible correlation between the two traits also constitutes evidence for discriminant validity (r = .10, p = .175).

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

DISCUSSION

The current study addressed whether the three online psycholinguistic measures tap the distinct construct from other time-pressured form-focused tests. Overall, the results of CFA confirmed that the two-factor model fit the data well (Hypothesis 1). Results of subset analysis demonstrated a different pattern for the two L2 groups varying in the amount of L2 experience (LOR). For the short-LOR group, the two-factor model did not converge, but the one-factor model produced a good fit. In contrast, the two-factor model, but not the one-factor model, fit the data well for the long-LOR group.

Construct validity of measures for automatized explicit and implicit knowledge

With regard to Hypothesis 2, although the factor loadings for automatized explicit knowledge were high and statistically significant (range = .65–.85), the loadings for implicit knowledge were much lower (range = .10–.48) in CFA. These low loadings underscore the challenges to devise measurements for implicit knowledge, indicating weak convergent validity for the measurements for implicit knowledge. Nevertheless, supporting evidence was provided for the discriminant validity, given a factor correlation below .80 (Brown, 2006; Hypotheses 3a and 3b). Although the one-factor model fit the data as well as the two-factor model, the substantial loadings in the one-factor model were all from the form-focused tasks. Moreover, the factor loadings from the three online measurements were all lower in the one-factor model than in the two-factor model. The MTMM analysis further showed stronger trait factors than method effects for both GJTs and reaction-time measures (Hypothesis 4).

Although a good deal of evidence has been provided for the construct validity of the hypothesized two-factor model, uncertainty is inevitably involved as to whether these two factors are automatized explicit and implicit knowledge. As proposed at the outset of this study (see Table 1 above), however, the tests for automatized explicit knowledge and implicit knowledge were designed to maximally differentiate the two types of tests in terms of the level of awareness involved during the test. Form-focused tasks such as timed GJTs and SPOT *directly* ask participants to pay attention to grammatical structures in stimulus sentences, raising the awareness of linguistic knowledge (i.e., explicit knowledge). Having said that, it is impossible to completely rule out the possibility that some L2 learners draw on implicit knowledge to perform timed GJTs. If learners were able to register the error online to make judgments in the timed GJT and had little reflection on their judgments, they might have relied primarily on implicit knowledge (Godfroid et al., 2015). With this inevitable ambiguous nature of GJTs in mind, however, if behavioral language tests are considered on a continuum spectrum from more explicit to more implicit, timed form-focused tasks like GJTs are probably considered closer to the explicit end of the continuum (DeKeyser, 2003; Vafaee et al., 2017).

In contrast, indirect real-time comprehension measures hypothesized to assess implicit knowledge never ask participants to detect errors. Instead, participants are asked to pay attention to the meaning of a sentence so that they can answer the comprehension question. This indirect nature of the grammar tests can prevent learners from becoming aware of their linguistic knowledge use and thus minimize the involvement of (automatized) explicit knowledge (Andringa & Curcic, 2015; Leung & Williams, 2012; Paradis, 2009; Suzuki & DeKeyser, 2015). Given these rationales of the test design, the current evidence suggests that the two factors should be labeled as automatized explicit knowledge and implicit knowledge. It casts doubt on the construct validity of the previous test battery of explicit and implicit knowledge developed by R. Ellis (2005) and further utilized by others (Bowles, 2011; Ercetin & Alptekin, 2013; Gutiérrez, 2013; Sarandi, 2015; Zhang, 2015). Although previous research is cautious in the interpretation that timed GJTs are a less pure measure for implicit knowledge (e.g., Loewen, 2009), time-pressure cannot guarantee the inaccessibility of automatized explicit knowledge (DeKeyser, 2003; Suzuki & DeKeyser, 2015; Vafaee et al., 2017).

The visual-world task is probably a superior measure to the RT measures because it requires no ungrammatical sentences, which makes the task most implicit. Furthermore, it directly captures real-time grammar processing via eye movements without any mediation such as through button presses. These advantages were empirically supported by the results from the CFA models (Figures 3 and 4) indicating that the factor loadings from the visual-world task were the highest in the current test battery. In contrast, the RT measures (self-paced reading and word-monitoring tasks) necessitate ungrammatical items, which may potentially raise awareness of the linguistic form. However, they are still useful assessment tools of implicit knowledge because L2 learners are unlikely to apply linguistic knowledge consciously when their grammatical processing is time locked within a few hundred milliseconds. When errors are registered without awareness, the level of noticing may sometimes rise up to consciousness as maintenance rehearsal is carried out in working memory. Regardless of the fact that registration may lead to further awareness *after the point of ungrammaticality*, implicit knowledge should be deployed for the registration of the errors *at the exact time of occurrence*, which is captured by the RT measures (Suzuki & DeKeyser, 2015). Furthermore, although the RT measures required button presses, the MTMM findings showed negligible method effects (Figure 5).

Further evidence: More L2 exposure leads to more stable use of implicit knowledge

There was a striking difference between the two LOR groups; the one-factor model produced a good fit for the short-LOR group, as opposed to the two-factor model for the long-LOR group (Hypothesis 1). Regarding Hypothesis 2, the factor loadings for the latent factor in the one-factor model (i.e., linguistic knowledge) suggest that L2 learners in the short-LOR group primarily rely on automatized explicit knowledge, as all the loadings for online measures were very low. Inspecting the results from the two-factor model in the long-LOR group, the factor loadings for automatized explicit knowledge were as good as for the whole group (range = .70–.80). It was critical that the factor loadings for implicit knowledge were higher and statistically significant: the two moderate loadings (.58 for the self-paced reading task and .62 for the visual-world task) and one weak loading (.39 for the word-monitoring task) in CFA.

The discriminant validity was further supported only for the long-LOR group (Hypotheses 3a and 3b), suggesting that both automatized explicit knowledge and implicit knowledge had been assessed more distinctively for them. In regard to the method effects (Hypothesis 4), the MTMM analysis for the long-LOR group further indicated that the correlated error of the two GJTs was significant but less than the trait factor loadings, and that of the word-monitoring task and the self-paced reading task was of nonsignificant small negative value. The traits seemed to be assessed more reliably with negligible method effects.

In sum, the overall findings from the long-LOR group supported all the hypotheses (except for the fit indices, Hypothesis 1, probably due to the smaller number of participants) more strongly, including the convergent validity of implicit knowledge. Even though implicit knowledge is much harder to assess, compared to automatized explicit knowledge, it is possible to tap into implicit knowledge with more stability, particularly when more experienced L2 learners performed the test battery. This corroborated the previous findings in Suzuki and DeKeyser (2015) and is consistent with Paradis's (2009) claim that explicit and implicit knowledge coexist in the L2 system, and the reliance of implicit knowledge increases over time through more L2 experience. Furthermore, regardless of the three analyzed groups, the factor loadings for automatized explicit knowledge were high (range = .63–.93). This suggests that late L2 learners with some formal instruction, as was the case for the present study, tend to rely on explicit knowledge very consistently (DeKeyser, 2007; Paradis, 2009). Results might be different when different L2 populations such as heritage learners and learners who received early foreign language instruction were administered with the current test set (e.g., Bowles, 2011; Phillip, 2009).

The caveat for the current subset analysis is that since each model consists of six indicators, even the rough estimation of the necessary sample size (10 participants $\times$ 6 indicators = 60) indicates that the sample size was less than ideal. The results from the subset analyses should be interpreted cautiously; however, it highlights a methodological gap in the previous studies. Most of the previous validation studies recruited classroom learners with limited immersion experience (Bowles, 2011; Gutiérrez, 2013; Zhang, 2015). The initial study by R. Ellis (2005, 2009) recruited L2 learners in the immersion context, but their LOR was relatively short (1.9 years). The present findings underscore that the amount of L2 exposure in the immersion context should be taken seriously for future research in the validation studies of measures of explicit and implicit knowledge.

Suggestions for further research

The current study opens several avenues for future research. More rigorous validation studies are needed for developing implicit knowledge measures. First, the reliability of the visual-world task was not assessed in the current study. Farris-Trimble and McMurray (2013) examined test–retest reliability of the visual-world task by requiring participants to complete the visual-world task for spoken word recognition twice (Day 1 and Day 2, which were separated by a week). The results showed that eye-movement patterns were closely related between Day 1 and 2, suggesting that the visual-world task is stable enough to index an individual's language processing. The present study could not assess the reliability of the task; it should be examined rigorously in future research.

Second, another point concerns the generalizability of the present findings. The current study focused on Japanese L2 learners experienced with both formal instruction and naturalistic environment. Further research is clearly needed in other L2 learner populations, different first language-second language combinations, other linguistic structures tested, and so on.

Third, the timed GJT tasks used here do not exactly follow the methodology of previous studies (e.g., R. Ellis, 2005; Zhang, 2015). In prior research, the time limit for each test item in timed GJTs was fixed to average response times by native speakers plus an extra 20% of the time, whereas the present study attempted to analyze the data with post hoc procedures. In the previous study where the time limit was imposed on the responses for each item on the test, test takers must have experienced more pressure and exhibited stronger motivation to make judgments quickly compared to the current format of GJTs. Further research should follow the exact design of the previous tasks in order to collectively advance the understandings in the previous measurements.

Conclusions

The present study set out to investigate the validity of more finely tuned implicit knowledge measures that are distinguishable from automatized explicit knowledge. An array of validity evidence supported that the six measurements assessed two distinct linguistic knowledge types, suggesting that indirect real-time comprehension tasks measured implicit knowledge, which was distinguished from automatized explicit knowledge. Although automatized explicit knowledge was assessed relatively easily by the conventional time-pressured form-focused tasks, it seems to be much harder to measure implicit knowledge through behavioral measures, particularly for L2 learners with less L2 exposure. Evidently, further investigations are needed as the theoretical distinction between automatized explicit knowledge and implicit knowledge can bear important implications for understanding SLA.

ACKNOWLEDGMENTS

This study was supported by IGERT: Biological and Computational Foundations of Language Diversity (NSF DGE-0801465), the Office of the Graduate Dean for a Summer Research Fellowship, the Language Learning Dissertation Grant Program, and the PhD program in Second Language Acquisition at the University of Maryland. This study is based on part of the author's doctoral dissertation, which was submitted to the University of Maryland College Park. I thank Robert DeKeyser, Yi Ting Huang, Steven Ross, and Nan Jiang for their advice on this project; Yuki Hirose and Edson Tadashi Miyamoto for their generous support with data collection; Kaoru Koyanagi, Yukiko Okuno, Tomomi Nishikawa, Hiromi Ozeki, and Kiyoko Tadokoro for their assistance in recruiting participants; and Kei Harata and Jun Fujita for their assistance in developing materials.

NOTES

1. The word *register* is used in the technical term in the current paper, meaning that cognitive registration of linguistic input that does not require awareness (see Suzuki & DeKeyser, 2015; Tomlin & Villa, 1994, for details).
2. Since this procedure was similar to the format of existing tests in the Japanese education system, where it is called the SPOT, this task is called the timed SPOT here (Kobayashi, Ford-Niwa, & Yamoto, 1996).
3. A third model was also constructed in terms of modality of measurements: a written-aural model. It never converged for the current data sets, however.
4. Japanese Language Proficiency Test (JLPT) N1 (which corresponds to the previous JLPT Levels 1) is roughly equivalent to the ACTFL Superior on the OPI scale (Kanno, Hasegawa, Ikeda, & Ito, 2005). JLPT Level 1 is the minimum requirement for acceptance into a regular college undergraduate/graduate program in Japan.
5. These two correlated errors were also imposed on the confirmatory factor analysis models; for parsimony, however, only the correlated errors that were statistically significant were retained in the final model.

6. Although the standardized root mean square and Benter–Bonnet nonnormed fit indices did not pass the criteria for the long-LOR group, they were close to the criteria. The overall assessment of the fit was deemed acceptable.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S014271641700011X

REFERENCES

Andringa, S., & Curcic, M. (2015). How explicit knowledge affects online L2 processing. *Studies in Second Language Acquisition, 37*, 237–268. doi:10.1017/S0272263115000017

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449–465.

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition, 33*, 247–271. doi:10.1017/S0272263110000756

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

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

DeKeyser, R. M. (1997). Beyond explicit rule learning. *Studies in Second Language Acquisition, 19*, 195–221.

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.

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

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: Routledge.

Elder, C., & Ellis, R. (2009). Implicit and explicit knowledge of an L2 and language proficiency. 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. 167–193). Tonawanda, NY: Multilingual Matters.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition, 27*, 305–352. doi:10.1017/S027226310505014X

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition, 27*, 141–172. doi:10.1017/S0272263105050096

Ellis, R. (2008). *The study of second language scquisition* (2nd ed.). Oxford: Oxford University Press.

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

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

Ercetin, G., & Alptekin, C. (2013). The explicit/implicit knowledge distinction and working memory: Implications for second-language reading comprehension. *Applied Psycholinguistics, 34*, 727–753. doi:10.1017/S0142716411000932

Farris-Trimble, A., & McMurray, B. (2013). Test–retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition. *Journal of Speech, Language, and Hearing Research, 56*, 1328–1345. doi:10.1044/1092-4388

Foote, R. (2011). Integrated knowledge of agreement in early and late English–Spanish bilinguals. *Applied Psycholinguistics, 32*, 187–220. doi:10.1017/S0142716410000342

Godfroid, A., Loewen, S., Jung, S., Park, J., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgements measure distinct types of knowledge. *Studies in Second Language Acquisition, 37*, 269–297. doi:10.1017/S0272263114000850

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning, 63*, 665–703. doi:10.1111/lang.12018

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research, 28*, 191–215. doi:10.1177/0267658312437990

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition, 35*, 423–449. doi:10.1017/S0272263113000041

Hasuike, I. (2004). Basho wo arawasu kakujoshi “ni” no kajou shiyou ni kansuru ichikousatsu: Chuukyuu reberu no chuugokuowasha no joshi sentaku sutoratejii bunseki [Investigation on the overuse of the particle “Ni” for location: Analysis of particle choice strategies by intermediate Chinese speakers]. *Nihongo Kyouiku*, *122*, 52–61.

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research, 29*, 33–56. doi:10.1177/0267658312461803

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

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. doi:10.1080/10705519909540118

Hulstijn, J. H. (2002). Towards a unified account of the representation, processing and acquisition of second language knowledge. *Second Language Research, 18*, 193–223. doi:10.1191/0267658302sr207oa

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning. *Studies in Second Language Acquisition, 27*, 129–140. doi:10.1017/S0272263105050084

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

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513–541. doi:10.1016/0749-596X(91)90025-F

Jiang, N. (2011). *Conducting reaction time research in second language studies*. New York: Routledge.

Jiang, N., Novokshanova, E., Masuda, K., & Wang, X. (2011). Morphological congruency and the acquisition of L2 morphemes. *Language Learning, 61*, 940–967. doi:10.1111/j.1467-9922.2010.00627.x

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

Kanno, K., Hasegawa, T., Ikeda, K., & Ito, Y. (2005). Linguistic profiles of heritage bilingual learners of Japanese. In J. Cohen, K. T. McAliser, K. Rolstad, & J. MacSwan (Eds.), *ISB4: Proceedings of the 4th International Symposium on Bilingualism* (pp. 1139–1151). Somerville, MA: Cascadilla Press.

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

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

Leung, J. H. C., & Williams, J. N. (2012). Constraints on implicit learning of grammatical form-meaning connections. *Language Learning, 62*, 634–662. doi:10.1111/j.1467-9922.2011.00637.x

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language, 63*, 447–464. doi:10.1016/j.jml.2010.07.003

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.

Loewenthal, K. M. (2004). *An introduction to psychological tests and scales* (2nd ed.). Hove: Psychology Press.

Matin, E., Shao, K., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics, 53*, 372–380. doi:10.3758/BF03206780

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

Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 69–89). Hillsdale, NJ: Erlbaum.

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

Philip, J. (2009). Pathways to proficiency: Learning experiences and attainment in implicit and explicit knowledge of English as a second language. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 194–215). Tonawanda, NY: Multilingual Matters.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control It. *Annual Review of Psychology, 63*, 539–569. doi:10.1146/annurev-psych-120710-100452

Posner, M. I., & Snyder, C. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing in cognition: The Loyola symposium* (pp. 55–85). Hillsdale, NJ: Erlbaum.

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*, 413–439. doi:10.1177/0267658313503171

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

Sedivy, J. C. (2010). Using eyetracking in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 115–138). Philadelphia: PA: Benjamins.

Spada, N. (2015). SLA research and L2 pedagogy: Misapplications and questions of relevance. *Language Teaching, 48*, 69–81. doi:10.1017/S026144481200050X

Suzuki, Y., & DeKeyser, R. M. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning, 65*, 860–895. doi:10.1111/lang.12138

Suzuki, Y., & DeKeyser, R. M. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*. Advance online publication.

Tanenhaus, M. K., & Trueswell, J. C. (2006). Eye movements and spoken language comprehension. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed.). New York: Elsevier.

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

Trenkic, D., Mirkovic, J., & Altmann, G. T. M. (2014). Real-time grammar processing by native and non-native speakers: Constructions unique to the second language. *Bilingualism: Language and Cognition, 17*, 237–257. doi:10.1017/S1366728913000321

Vafaee, P., Suzuki, Y., & Kachinske, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition, 39*, 59–95. doi:10.1017/S0272263115000455

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1–26. doi:10.1177/014662168500900101

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). London: Emerald Group.

Zhang, R. (2015). Measuring university-level L2 learners' implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition, 37*, 457–486. doi:10.1017/S0272263114000370