In press in *Studies in Second Language Acquisition* (Cambridge University Press)

# Conflicting Evidence of Explicit and Implicit Knowledge from Objective and Subjective Measures

Ryo Maie
Michigan State University

Robert M. DeKeyser
University of Maryland

## Abstract

This study is the first to compare objective and subjective measures of explicit and implicit knowledge under learning from incidental exposure. An experiment was conducted, during which L1 English speakers were trained on a semi-artificial language, Japlish. A measure of explicit knowledge and a recently proposed measure of implicit knowledge (i.e., an untimed auditory grammaticality judgment and a word-monitoring task) were applied to gauge the two types of knowledge at two testing sessions, and their results were compared with those of subjective measures of awareness. Results revealed clear discrepancies between the two measurement approaches in terms of their sensitivity. In particular, while the subjective measures varied in identifying explicit and implicit knowledge of various Japlish constructions, the objective measures indicated that most of the knowledge was explicit, and development of implicit knowledge (measured by the word-monitoring task) was minimal, only manifested in detecting a case-missing violation at the delayed posttest. The results are discussed with reference to the current literature on explicit and implicit learning and knowledge, and it is concluded that the criterion of (un)awareness might not be by itself sufficient to provide a full account of L2 knowledge developed under incidental conditions.

## Acknowledgement

We are grateful to two anonymous reviewers as well as the journal editor, Dr. Susan Gass, for their constructive feedback on the earlier drafts of the paper. This article is based on the original master's thesis of the first author, supported by the Program in Second Language Acquisition at the University of Maryland, College Park. We would like to express our sincere gratitude to Drs. Michael Long and Steven Ross for their insightful suggestions on the thesis and Dr. Yuichi Suzuki, Dr. Amelia Lambelet, and Wei Yi for their comments on the earlier drafts of the paper. The earlier version of this piece was presented at Second Language Research Forum 2018 at Université du Québec à Montréal. All errors remain our own and any questions regarding the paper should be addressed to the corresponding author.

Conflicting Evidence of Explicit and Implicit Knowledge
from Objective and Subjective Measures

A long line of L2 research has corroborated that adult language learning entails both explicit and implicit processes (see Rebuschat, 2015 for a review). Explicit learning is a conscious operation in which learners are aware of what is being learned, whereas implicit learning takes place incidentally without awareness. They are each argued to result in two qualitatively distinct types of L2 knowledge, namely, explicit and implicit knowledge (EK and IK), with the former being knowledge one is aware of (therefore *often* consciously accessible and reportable), while the latter being tacit knowledge upon which one is not able to retrospect (DeKeyser, 2009). Recent SLA research has suggested that adults can develop both EK and IK from brief incidental exposure to the target language (e.g., Grey, Williams, & Rebuschat, 2014; Kachinske, Osthus, Solovyeva, & Long, 2015; Rebuschat & Williams, 2012). However, the literature is still characterized by diverse methodological approaches presenting contrasting results. While Williams (2005), for instance, adopted offline (or retrospective) verbal reports to gauge awareness of target structures, Hama and Leow (2010) utilized a triangulation of both offline and concurrent verbal reports. The two present conflicting results regarding the plausibility of implicit learning by adults, and yet another group of studies reports a different picture, utilizing subjective measures of awareness (see below).

This study presents a novel perspective on the issue by incorporating into the literature objective measures of EK and IK recently proposed in SLA research (e.g., Suzuki, 2017; Vafaee, Suzuki, & Kachinske, 2017). An experiment was conducted, in which participants were trained on a semi-artificial language under incidental conditions.[1] The product of learning was assessed by two objective measures of EK and IK (i.e., an untimed auditory grammaticality judgment task and a word-monitoring task) and their results were compared with those of subjective measures of awareness.

## Explicit and Implicit Knowledge under Incidental Conditions

### Explicit and Implicit Knowledge Assessed with Verbal Reports

Williams (2005) is widely considered to have initiated the current line of research on explicit and implicit learning under incidental conditions (though its primary focus was on demonstrating implicit learning by adults). The study showed that after incidentally exposed to sentences containing target determiner-noun constructions, participants were significantly above chance on choosing the correct combinations despite being unaware of the rule investigated. Awareness of the rule was gauged through retrospective offline verbal reports at the end of the experiment, and it was assumed that if the participants were not able to verbalize the rule, they were unaware of it, which, therefore, indicated development of IK. Hama and Leow (2010) later criticized the study for making such unwarranted assumption, and instead argued that awareness

need be examined concurrently at the stage of knowledge encoding rather than retrospectively at the stage of knowledge retrieval.

Methodologically speaking, verbal reports suffer from the problem of memory decay by participants such that memory of conscious processing episodes may have decayed at the time when participants are asked to report on them. Moreover, the absence of verbalization does not necessarily indicate the absence of awareness because participants can be aware of rules they have learned, but they might not be able to verbalize because they cannot explain verbally, or they are not confident enough to talk about them (Rebuschat, 2013; Shanks & St. John, 1994). Hama and Leow (2010) did not find evidence of implicit learning when the awareness was assessed through a combination of both retrospective and concurrent verbal reports. As a result, there now seems to be a consensus among researchers that retrospective reports can (but not necessarily will) be an insensitive measure of awareness (Rebuschat, 2013; Shanks & St. John, 1994) and it is probably so when the rule is complex and abstract in nature. Moreover, researchers agree that a triangulation of multiple measures is required to understand the complex and multidimensional nature of L2 knowledge in more detail (e.g., Chan & Leung, 2018; Rebuschat et al., 2015). Note, however, that the results of Hama and Leow (2010) were also criticized on the ground that the use of concurrent reports is associated with its own problem; namely that the act of verbalization may change the underlying cognitive processes being investigated (i.e., the issue of reactivity, see Rebuschat, et al., 2015), therefore obliterating implicit processes that otherwise would have happened.

**Explicit and Implicit Knowledge Assessed with Subjective Measures of Awareness**

To overcome the issue on the use of the verbal reports, recent studies have begun to adopt subjective measures of awareness originating from research in cognitive psychology (see Rebuschat, 2013 for a review). Subjective measures here refer to confidence ratings and source attributions that are normally paired with a judgment task (e.g., grammaticality judgment tasks), during which participants are asked to indicate their confidence in, and the perceived source of, each judgment they make. The proposition underlying the use of subjective measures is that knowledge is unconscious when one objectively performs above chance but does not know subjectively that he/she actually does (Cheesman & Merikle, 1984). Dienes and his collaborators (e.g., Dienes, 2008; Dienes, Altmann, Kwan, Goode, 1995) have advocated that awareness be measured in reference to one's subjective account of his/her performance and use two criteria as requirements that must be satisfied as evidence of IK. *The guess criterion* can be met when one performs above chance despite the fact that he/she claims to be guessing (obtained from both confidence ratings and source attributions), and the zero-correlation criterion can be met when there is no correlation between accuracy of one's performance and the subjective level of confidence he/she reports (based on confidence ratings). It is argued that the subjective measures are more sensitive to low-level confidence that is not often detected by retrospective verbal reports, because they do not require participants of verbalization (Rebuschat & Williams, 2012; Rebuschat et al., 2015). Indeed, existing L2 studies with subjective measures have found that conscious awareness is facilitative of learning under incidental conditions (therefore, EK), but along with it

participants also develop knowledge that they are not aware of (e.g., Grey et al., 2014; Kachinske et al., 2015; Rebuschat & Williams, 2012).

Subjective measures, however, are not without problems either. They can suffer from what is called 'response bias', which refers to a phenomenon that participants may systematically claim that they are guessing (or not confident at all) even though they in fact have some degree of awareness (Dienes, 2008; Rebuschat, 2013). For instance, participants often set their own criteria on rating their level of confidence, and those who are conservative may not report that they are confident unless they are absolutely certain. Kunimoto, Miller, and Pashler (2001) suggested that this can be resolved statistically with use of the *d'* (d-prime) statistic because it is computed by accounting for both sensitivity and noise (i.e., hit and false alarm) in participants' responses.

More importantly, however, there is now a reason to believe that the practice of pairing a judgment task with subjective measures may not be a felicitous way to measure IK. As discussed below, it has been shown that tasks such as grammaticality judgment tasks (GJTs) do not tap into implicit L2 knowledge even when applied in timed conditions (e.g., Suzuki, 2017; Vafaee et al., 2017) because instructions as well as the nature of the tasks necessarily orient participants to focus on accuracy and form of the language in question, which can dispose participants to consciously reflect upon their judgment processes (Suzuki & DeKeyser, 2017). Although one may argue that objective measures are not process-pure tasks (see Rebuschat, 2013 for a review), it is clear that participants draw on EK to a much larger extent than on IK when they perform GJTs. It thus seems rather self-contradictory to examine whether participants draw on IK when they perform an EK measure such as GJTs.

**Recent Advancements in Measurement of Explicit and Implicit Knowledge**

Fortunately, recent validation studies of objective EK and IK measures have investigated new measures of IK (e.g., a word-monitoring and self-paced reading task) that can broaden our perspective of what can be counted as evidence of IK. For more than a decade, SLA researchers have sought language tasks that can reliably gauge EK and IK of L2. R. Ellis (2005) conducted the seminal study that carried out a psychometric analysis of five often-used language tasks. Based on a principal component analysis, results revealed that the tasks converged on two latent factors interpreted as EK and IK. More importantly, tasks including elicited imitation, oral narratives, and timed GJT loaded onto the factor of IK, whereas others such as untimed GJT and metalinguistic knowledge test loaded onto the other, EK. Some researchers, however, recently conducted follow-up studies that questioned the validity of the conclusions made by R. Ellis. They argue that the tasks that were concluded to be a measure of IK (i.e., elicited imitation and timed GJT) do not in fact tap into IK, but rather, measure EK whose use and access have been automatized (Suzuki, 2017; Vafaee et al., 2017).

Through consistent practices, L2 learners can automatize the use of EK, developing automatized EK (DeKeyser, 2015). It comes as no surprise, therefore, that participants can perform tasks such as timed GJTs or elicited imitation with EK because they cannot necessarily limit the use of it when it is accessed rapidly (DeKeyser, 2003). However, as automatization is a long and

slow process to complete (i.e., reaching asymptote), learners often end up with partially automatized knowledge, and may never achieve its full automatization. While fully automatized EK is conceptualized to be functionally equivalent to IK in that its access can be fast and occur without conscious awareness, partially automatized EK is still distinguishable from IK by the criterion of awareness (DeKeyser, 2017; Suzuki & DeKeyser, 2017 for discussions).

To validate the line of claims above, Suzuki and collaborators proposed new measures of IK, and compared them against the putative IK measures identified by R. Ellis (2005). Suzuki (2017), for instance, conducted a multitrait-multimethod analysis of six measures of L2 knowledge: an eye-tracking visual-world paradigm (Eye), a word-monitoring task (WMT), a self-paced reading task (SPR), a timed auditory and visual GJT (T-AGJT; T-VGJT), and a timed SPOT task (Simple Performance-Oriented Test: T-SPOT).[2] Results of two confirmatory factor analyses demonstrated that for those with a longer length of residence who were expected to have developed both EK and IK, a two-factor model that included IK and automatized EK better explained the participants' performance than a one-factor model that contained only one construct, language knowledge. The study found that Eye, WMT, and SPR loaded onto the same factor labeled as IK, whereas T-AGJT, T-VGJT, and T-SPOT loaded onto the other, automatized EK. Suzuki concluded that the online RT-based measures are the ones that can reliably gauge IK, whereas T-AGJT, T-VGJT, and elicited imitation "should be deemed a measure of automatized explicit knowledge" (p. 1231). A recent study by Godfroid, Kim, Hui, and Isbell (2018) conducted an analysis of a yet larger number of linguistic outcome tasks ($k$ = 12) that directly tested a two-factor solution with EK and IK against a three-factor solution with EK, automatized EK, and IK. Their model comparison revealed that neither of them were significantly better than the other, but a close examination of factor loadings suggested that the three-factor solution may have an advantage.

In sum, recent studies of EK and IK measures provided strong support for the new objective measures. A question should be asked then as to what extent these measures contribute to the lines of evidence on development of EK and IK under incidental conditions reviewed above. For instance, do they differ from subjective measures of awareness in the extent to which they identify participants having developed EK and/or IK from brief incidental exposure? To answer this question, the present study adopted a word-monitoring task as an objective measure of IK, and its results were compared with those of subjective measures. Although the study was novel in employing a word-monitoring task, it was conducted in the same spirit of the previous research that triangulated multiple measures of EK and IK to better understand the complex nature of developing L2 system (e.g., Chan & Williams, 2018; Rebuschat et al., 2015). Indeed, there is an acute need for experimental studies that further triangulate products of learning under incidental conditions, as individual studies markedly diverge as to whether implicit learning is viable for adult learners and what types of L2 knowledge they develop as a result. For instance, several studies claim to have shown evidence that abstract form-meaning mappings can be learned implicitly (e.g., Chan & Leung, 2018; Leung & Williams, 2011, 2012; Williams, 2005), whereas others found that implicit learning was limited to learning of form-form associations (e.g.,

DeKeyser, 1995; Godfroid, 2016). A similar debate has taken place in cognitive psychology as well (see Knowlton & Squire, 1994; Perruchet & Pacteau, 1990 for examples).

## The Study

The study consisted of a laboratory experiment wherein participants were exposed to a semi-artificial language, *Japlish*. They were tested at two sessions, once immediately after the training and again two weeks later. This was to reflect results of recent incidental learning studies that effects of incidental exposure may not surface until some period after the exposure (e.g., Grey et al., 2014; Morgan-Short, Finger, Grey, & Ullman, 2012). Two objective measures assessed the learning, an untimed auditory grammaticality judgment task (U-AGJT) and a word-monitoring task (WMT). The former was employed as a measure of EK that allowed for controlled processing of L2 knowledge, whereas the latter was that of IK that required the participants to use automatic, and possibly, implicit processing of L2 knowledge. The participants performed the subjective measures while they made grammaticality judgments. Based on the research design, three main research questions were addressed:

1. Do adult L2 learners develop EK and/or IK from brief incidental exposure measured by two objective outcome measures, U-AGJT and WMT?
2. To what extent do results from U-AGJT and WMT converge on or diverge from those of subjective measures of awareness?
3. Do the findings for research question 1 and 2 change after two weeks of delay with no exposure?

## Method

### Participants

Participants were 63 L1 speakers of English ($M_{age}$ = 19.47, $SD_{age}$ = 1.78, *Min* = 18, *Max* = 27) who were all undergraduate or graduate students at a state university on the East Coast of the United States. At the time of the experiment, no one had experience with Japanese nor any case-marking languages and none had stayed in countries where a case-marking language was spoken for more than two weeks. Data of 14 participants were excluded from this sample because they did not return for the delayed posttest ($n$ = 8), did not follow the instructions ($n$ = 4), or produced a mean WMT latency larger than 2500 ms ($n$ = 2), resulting in a final sample of 49 participants (38 females). Although their L1 was English, they also knew one or more second languages ($M$ = 1.29, $SD$ = 0.61, *Min* = 0, *Max* = 3) at various proficiency levels (beginner to advanced). The participants were randomly assigned to either the experimental group ($n$ = 28) or the control group ($n$ = 21) at the beginning of the experiment. While the experimental group was exposed to exemplar sentences with fixed patterns, the control group listened to sentences whose word order and the position of case markers were pseudo-randomized so that the participants would be exposed to every possible word-order pattern with the same frequency.

**Materials**

**Exposure task material**. Japlish was originally used by Williams and Kuribara (2008) in their study of learning under incidental conditions. In the present experiment, the language consisted of four word orders and three case markers of Japanese. There were two simple (OSV and OSIV) and two complex (OSSVV and OSSIVV) word orders, depending on how many clauses each sentence contained.[3] The three case markers conveyed different grammatical information, with *-ga* for the subject, *-o* for the direct object, and *-ni* for the indirect object. An example sentence is provided for each word order type below:

a. O-S-V
   *That wall-o Mary-ga painted*
   "Mary painted that wall"
b. O-S-I-V
   *The picture-o John-ga his friend-ni sent*
   "John sent the picture to his friend"
c. O-S-[S-V]-V
   *The tuition-o Mary-ga her school-ga raised said*
   "Mary said that her school raised the tuition"
d. O-S-[S-I-V]-V
   *Those documents-o John-ga his workmate-ga their boss-ni faxed said*
   "John said that his workmate sent those documents to their boss"

A total of 100 sentences were constructed and checked with L1 speakers of American English on whether they made sense or not. Out of the entire set, 25 sentences corresponded to each word order type, each of which contained thirteen semantically plausible and twelve implausible sentences. The entire set was presented twice to the participants, amounting to 200 trials in total. The order of the presentation was randomized regardless of the complexity of each sentence, following Grey et al. (2014). For each word order type (i.e., OSV, OSIV, OSSIVV, and OSSIVV), the sentence length was controlled as four, six, seven, and nine words, respectively. See Appendix A for the entire list of the exposure stimuli (Online Supplementary Material).

**Untimed auditory grammaticality judgment task**. Eighty Japlish sentences were constructed, 32 grammatical and 48 ungrammatical sentences (8 items for each grammatical and ungrammatical item type). Ungrammatical items were devised such that they had an illicit word order, contained a noun whose case marker was missing, or had the positions of two case markers reversed. Table 1 illustrates ungrammatical items introduced in U-AGJT. For items with a case-marking violation, sentences were presented in OSIV for two reasons. First, this word order type contained all case markers in question (Grey et al., 2014), and second, the control of the type of word order allowed for a reliable measurement of knowledge of the case markings alone, without confounding it with that of the word orders. Four lists of presentation were created by first writing two lists of items with entirely different sentences and counterbalancing them by grammaticality.

The participants were presented with two of the presentation lists at the immediate and the delayed posttest, and care was taken so that they would not see the same sentence twice. Appendix B (supplementary materials) presents the entire list of the U-AGJT items.

Table 1. *Ungrammatical Item Types in Testing Tasks*

| | Word Order |
|---|---|
| O-V-S | *The fire-o lighted Angela-ga* |
| O-S-V-I | *The letter-o Mary-ga faxed her boss-ni* |
| O-S-V-S-V | *That vase-o Stacey-ga broke her spouse-ga thought* |
| O-S-S-V-I-V | *The document-o Mary-ga her workmate-ga faxed her boss-ni that told* |
| | Case Missing |
| -ga | *A tip-o Tim the driver-ni gave* |
| -o | *The clothes Pamela-ga her daughter-ni chose* |
| -ni | *A letter-o Steve-ga the mayor wrote* |
| | Case Mixing |
| -ga, -o | *A tip-ga Tim-o the driver-ni gave* |
| -ga, -ni | *The clothes-o Pamela-ni her daughter-ga chose* |
| -o, -ni | *A letter-ni Steve-ga the mayor-o wrote* |

During the task, the participants judged whether each sentence conformed to the patterns of the language by pressing corresponding YES/NO keys. Although they were only played once, it was expected that the untimed nature of the task allowed the participants to consciously reflect on their judgments and draw more on EK they developed from the exposure (Suzuki & DeKeyser, 2017; R. Ellis, 2005; Vafaee et al., 2017). Furthermore, the participants also indicated the level of confidence on each response and what source of knowledge they relied upon. For the confidence ratings, the participants used a scale of 1–5 key on a keyboard, which corresponded to each confidence level, 1 = "guess: 50%", 2 = "somewhat confident: 60-70%", 3 = "confident: 70-80%", 4 = "very confident: 80-90%", and 5 = "absolutely certain: 100%".[4] For the source attributions, keys 1–4 were assigned, each of which corresponded to 1 = "guess", 2 = "intuition", 3 = "memory of items from the exposure phase", and 4 = "rule". The granularity of the scales was adopted from Kachinske et al. (2015) with slightly different labels (see Rebuschat, 2013 for guidelines). The meaning of each confidence level and source attribution category was carefully explained; for instance, participants were instructed to select "guess: 50%" (confidence) and "guess" (source) category when their response was based on a complete guess (i.e., 50/50), and "intuition" when they felt the sentence was grammatical (or ungrammatical) but they did not know why.[5]

**Word-monitoring task**. 130 Japlish sentences were constructed, 96 of which were target sentences. Of the targets, half were grammatical and the other half were ungrammatical (8 items for each grammatical and ungrammatical item type). All ungrammatical item types were the same as those for U-AGJT (see Table 1), and four lists of presentation were also created in the same manner. However, entirely new sentences with the same structures were written for WMT, so as to ensure that the participants would never hear the same sentence twice. Furthermore, the items also contained a multi-word adverb phrase at the beginning and the end of a sentence. This was to ensure that ungrammatical elements would not come at the place of the first or the second word,

nor the last or the second-to-last word of the sentence because as pointed out by Jiang (2012), participants tend to be less focused at the beginning or the end of a sentence than they are in the middle. Each target word to be monitored was chosen such that it immediately followed the ungrammatical element in the sentence (see Appendix C for details).

During the task, the participants first saw an asterisk for 500 ms that subsequently turned into a target word to monitor. The target word was presented visually and remained on the screen while a recording of a carrier sentence was played. The participants were told to focus on the meaning of the sentence and the appearance of the target word, and press a designated key as soon as they heard the target. With a probability of one in two sentences, a comprehension question followed in order to ensure that the participants would focus on the meaning of each sentence. This dual-task condition minimized the possibility that they would use any EK of grammar or conscious strategies. Appendix C lists the WMT stimuli.

**Procedure**

The study was introduced as a project wherein the experimenter investigated whether native speakers of English were able to comprehend an artificial language that had been developed recently. Although the participants were notified that a few tests would follow the training phase, they were told that the tests would be on comprehension of sentences, not on grammar. At this point, each participant was randomly assigned to either the experimental or the control group.
During the exposure phase, the participants performed semantic plausibility judgments. Using two contrasting sentences in English (i.e., *John ate an apple* vs. *John ate a chair*), it was described that their task was to judge the plausibility of each sentence in meaning. They were informed that the language they would listen to was not English, but they would be able to understand it. Immediately after the exposure task, they were told that the study was actually on learning of the artificial language, not comprehension, and they would be subsequently tested on the knowledge they had acquired. WMT immediately followed as the first testing task, described as a test of language processing and comprehension so as not to disclose the existence of ungrammatical items. For U-AGJT, the participants were instructed to make a judgment on whether each sentence was grammatical or not, based on their experience in the exposure phase. Additionally, they were also asked to perform the confidence ratings and source attributions on each judgment. As suggested by Hamrick and Sachs (2018), the instructions throughout the study were kept the same for both groups. DMDX software was used to conduct the entire experiment (Forster & Forster, 2003), and each testing task had eight practice items before the test items.

Two weeks later, the participants came back to the laboratory and began the second session with WMT and then U-AGJT. They also took a post-experimental questionnaire, in which they were asked to perform offline verbal reports on their noticing and understanding about the language (Appendix D for the questionnaire). Due to space constraints, however, results on the verbal reports are not reported here but summarized in Appendix I. Upon completing the questionnaire, the participants were thanked for their participation and received compensation.

**Analysis**

To account for potential response biases, the main analysis of U-AGJT data was carried out using the d-prime statistic (Kunimoto et al., 2001). The d-prime score is a sensitivity index corrected for one's response bias. 0 represents complete inability to discriminate (i.e., a chance-level performance), and the score can range from 4.65 (the effective limit: 99%) to –4.65 (1%) with a higher score representing better discriminatory ability. Any d-prime scores that exceeded +/– 3SD of the corresponding group's average were excluded from the analysis. Group differences on the d-prime scores were first examined through a multivariate analysis of variance (MANOVA) with Group as a between-subjects factor (two levels: Experimental and Control) and d-prime scores on six construction types as dependent variables (i.e., OSV, OSIV, OSSVV, OSSIVV, CaseMis, and CaseMix). If the multivariate analysis detected any significant main effect of Group, follow-up univariate analyses of variance (ANOVAs) were conducted to examine on which dependent variables the groups significantly differed from each other.

To test whether the guessing and zero-correlation criteria were satisfied, two steps of analysis were taken to follow the current practice in the literature. First, the Chan difference score (i.e., the difference in confidence between correct and incorrect judgments) was obtained for each construction type by calculating mean confidence levels on correct and incorrect responses and examining whether they differed significantly from each other (Chan, 1992; Rebuschat, 2013). Any significant difference in which the confidence level on the correct responses was higher than that on the incorrect responses was interpreted as evidence that the participants were aware of knowledge they had developed. Second, mean accuracy on each source attribution category was calculated to see if the participants performed significantly above chance when they claimed to be drawing on a given source of knowledge. Any performance that was significantly above chance when the participants claimed to be guessing or drawing on intuitions was taken as evidence of IK (but see endnote 5 for a caveat).

For the analysis of WMT data, mean raw reaction times (RTs) on overall, grammatical, and ungrammatical items were calculated across different construction types to interpret the data descriptively. Before summarizing, any item-level RTs that exceeded +/– 3SD of the person's mean and that were larger than 2500 ms and smaller than 100 ms were excluded from the analysis (cf. Jiang, 2012). This affected 4.08% (immediate) and 4.18% (delayed) of the entire dataset. Furthermore, individual means that exceeded +/– 3SD of the corresponding group's average were also excluded, resulting in an exclusion of two participant means. The mean RTs were transformed into their reciprocals and multiplied by –1000 (TransRTs) so as to reduce the positively skewing nature of RT data. As with the U-AGJT data, group differences on TransRTs were first examined through a MANOVA with Group as a between-subjects factor, with Grammaticality as a within-subjects factor, and with TransRTs on each construction type as dependent variables. Univariate ANOVAs were conducted to see on which construction types the experimental group and the control group significantly differed from each other in their word-monitoring latencies to grammatical and ungrammatical items. A significant main effect of Grammaticality for which RTs

on ungrammatical items were larger than those on grammatical items was interpreted as evidence of implicit sensitivity to grammatical violation.

Before the main analyses, assumptions associated with each statistical test were examined, and care was taken to make sure that any violation of statistical assumptions would not make a significant impact on the analyses. See Appendix E for results. Furthermore, due to technological difficulties, six d-prime scores from one participant for Immediate U-AGJT were missing. Since the other data points including d-prime scores on Delayed U-AGJT, TransRTs on Immediate and Delayed WMT, and scores on four cognitive aptitudes (reported elsewhere) were available for that individual, those data points were imputed using the *R* package *mice* (version 2.46.0, van Buuren & Groothuis-Oudshoorn, 2011).

## Results

### Untimed Auditory Grammaticality Judgment Task

**Immediate posttest**. The internal consistency of the task based on Cronbach's alpha was $\alpha = 0.92$. Overall mean accuracy for the experimental and the control groups were 73.65% (*SD* = 27.76) and 56.48% (*SD* = 32.82), respectively (see Table S13 and S14 in Appendix F). In fact, the experimental group outperformed the control group in all respects, except the mean accuracy for the grammatical items of OSV sentences (the experimental, 84.35% and the control, 88.33%). A MANOVA on d-prime scores showed that there was a significant main effect of Group at the multivariate level, $F(1, 47) = 3.53$, $p < .006$; Pillai's trace V = .34. Subsequent follow-up ANOVAs showed that the experimental group outperformed the control group on all of the construction types with mostly large effect sizes, according to the benchmark suggested in Kirk (1996): $\omega^2 = .010$ (small), .059 (medium), and .138 (large); $F(1, 47) = 12.64$, $p < .001$, $\omega^2 = .19$ for OSV; $F(1, 47) = 9.29$, $p = .003$, $\omega^2 = .14$ for OSIV; $F(1, 47) = 15.99$, $p < .001$, $\omega^2 = .23$ for OSSVV; $F(1, 42) = 12.82$, $p < .001$, $\omega^2 = .19$ for OSSIVV; $F(1, 47) = 8.65$, $p = .005$, $\omega^2 = .13$ for Case Missing; $F(1, 42) = 5.46$, $p = .024$, $\omega^2 = .08$ for Case Mixing. See Figure 1 for a graphical summary.

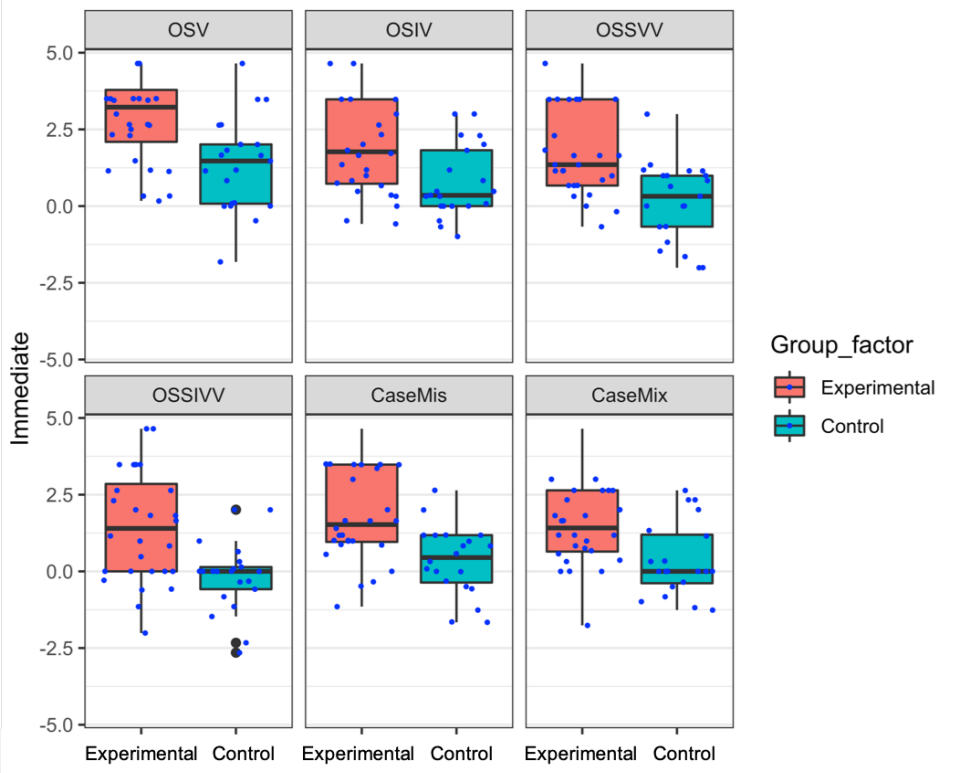


FIGURE 1. Group differences on immediate U-AGJT.
*Note*. Red = Experimental and Blue = Control

**Delayed posttest**. The internal consistency of the task was α = 0.94. After two weeks with no exposure, the experimental group still outperformed the control group with mean accuracy of 69.12% (*SD* = 19.17) and 50.18% (*SD* = 16.02), respectively (see Table S15 and S16 in Appendix F). Again, the experimental group outperformed the control group in all respects except the mean percentile accuracy for grammatical OSV items (the experimental, 85.71%, and the control, 93.45%). This was due to the fact that both groups were more inclined to accept sentences rather than to reject them, and the bias was especially strong for the control group. Recall that sentences that the control group was exposed to had word orders and positions of case markers pseudo-randomized such that any word orders and positions of case markers were possible for them. They were thus expected to consider the language as random and this probably made it more likely that they would accept sentences because of less stable knowledge they could have developed from the exposure (if any).

Results of a MANOVA on d-prime scores mirrored those of the immediate posttest, indicating that there was a significant main effect of Group at the multivariate level, $F(1, 46) = 2.54$, $p < .034$; Pillai's trace V = .27. Furthermore, follow-up ANOVAs showed that the experimental group outperformed the control group on all of the construction types, $F(1, 46) = 7.64$, $p = .008$, $\omega^2 = .12$ for OSV; $F(1, 46) = 4.25$, $p = .044$, $\omega^2 = .06$ for OSIV; $F(1, 46) = 7.96$, $p = .007$, $\omega^2 = .12$ for OSSVV; $F(1, 46) = 10.69$, $p = .002$, $\omega^2 = .16$ for OSSIVV; $F(1, 46) = 10.74$, $p = .002$, $\omega^2 = .16$ for Case Missing; $F(1, 46) = 4.92$, $p = .031$, $\omega^2 = .07$ for Case Mixing. Although

the effect seemed variable across the construction types, it was particularly robust for the complex word orders. Figure 2 graphically summarizes the comparison of the experimental and the control groups at the delayed posttest.

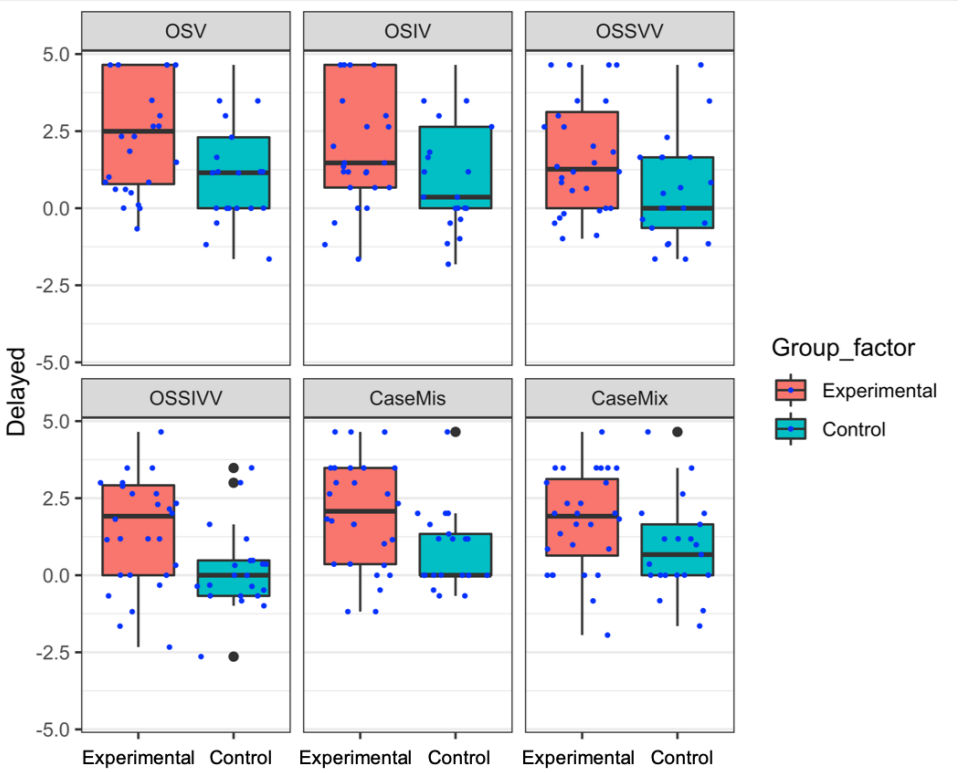


FIGURE 2. Group differences on delayed U-AGJT.
*Note*. Red = Experimental and Blue = Control

In order to examine differences between the experimental group's scores at the immediate and the delayed posttest, a repeated-measures two-way ANOVA was conducted for the experimental group only, with Construction and Time (i.e., immediate vs. delayed) as within-subjects factors. Note that Mauchly Tests for sphericity reached significance for Construction and the interaction of Construction and Time, $\chi^2(5) = 0.35$, $p = .024$, and $\chi^2(5) = 0.36$, $p = .031$, respectively. Therefore, the Greenhouse-Geisser correction was made to the corresponding degrees of freedom. Results revealed that there was a significant main effect of Construction, $F(5, 135) = 8.20$, $p < .001$, $\omega^2 = .06$, but no main effect of Time, $F(1, 27) = 0.0006$, $p = .97$, $\omega^2 = .000$, or interaction of Construction and Time, $F(5, 135) = 0.88$, $p = .478$, , $\omega^2 = .003$, indicating that the experimental group performed as well at the delayed posttest as at the immediate posttest.

**Subjective Measures of Awareness**

**Immediate posttest**. Results of the confidence ratings and the source attributions at the immediate posttest are summarized in Table 2 and Figure 3, respectively. The confidence level on correct responses was significantly higher than that on incorrect responses for all of the word order

types, and this was still the case even after Bonferroni corrections, except for Case Missing. Based on the effect sizes, however, the association between awareness and the accuracy of performance seemed small, as none of the differences were larger than what was argued to be a small effect size in Plonsky and Oswald (2014), $d = 0.60$. Lastly, there was no significant difference in the confidence level on Case Mixing, and the participants in fact performed significantly above chance on this construction type, $t(27) = 5.80$, $p < .001$, 95% CI [1.00, 2.09], suggesting a beginning development of IK.

Table 2. *Mean Confidence Levels on Immediate U-AGJT*

| Construction | Correct (*SD*) | Incorrect (*SD*) | *t* | *df* | *SE* | Cohen's *d* |
|---|---|---|---|---|---|---|
| OSV | 3.87 (1.18) | 3.31 (1.26) | 5.31** | 282 | 0.10 | 0.37 |
| OSIV | 3.47 (1.20) | 3.07 (1.24) | 4.29** | 469 | 0.09 | 0.34 |
| OSSVV | 3.15 (1.20) | 2.78 (1.03) | 4.50** | 590 | 0.09 | 0.33 |
| OSSIVV | 3.19 (1.27) | 2.71 (1.11) | 5.45** | 720 | 0.11 | 0.39 |
| CaseMis | 3.47 (1.22) | 3.21 (1.13) | 2.11* | 173 | 0.11 | 0.21 |
| CaseMix | 3.04 (1.28) | 3.27 (1.15) | -1.75 | 321 | 0.12 | 0.15 |

*Note*. * $p < .05$. ** significant after Bonferroni correction.

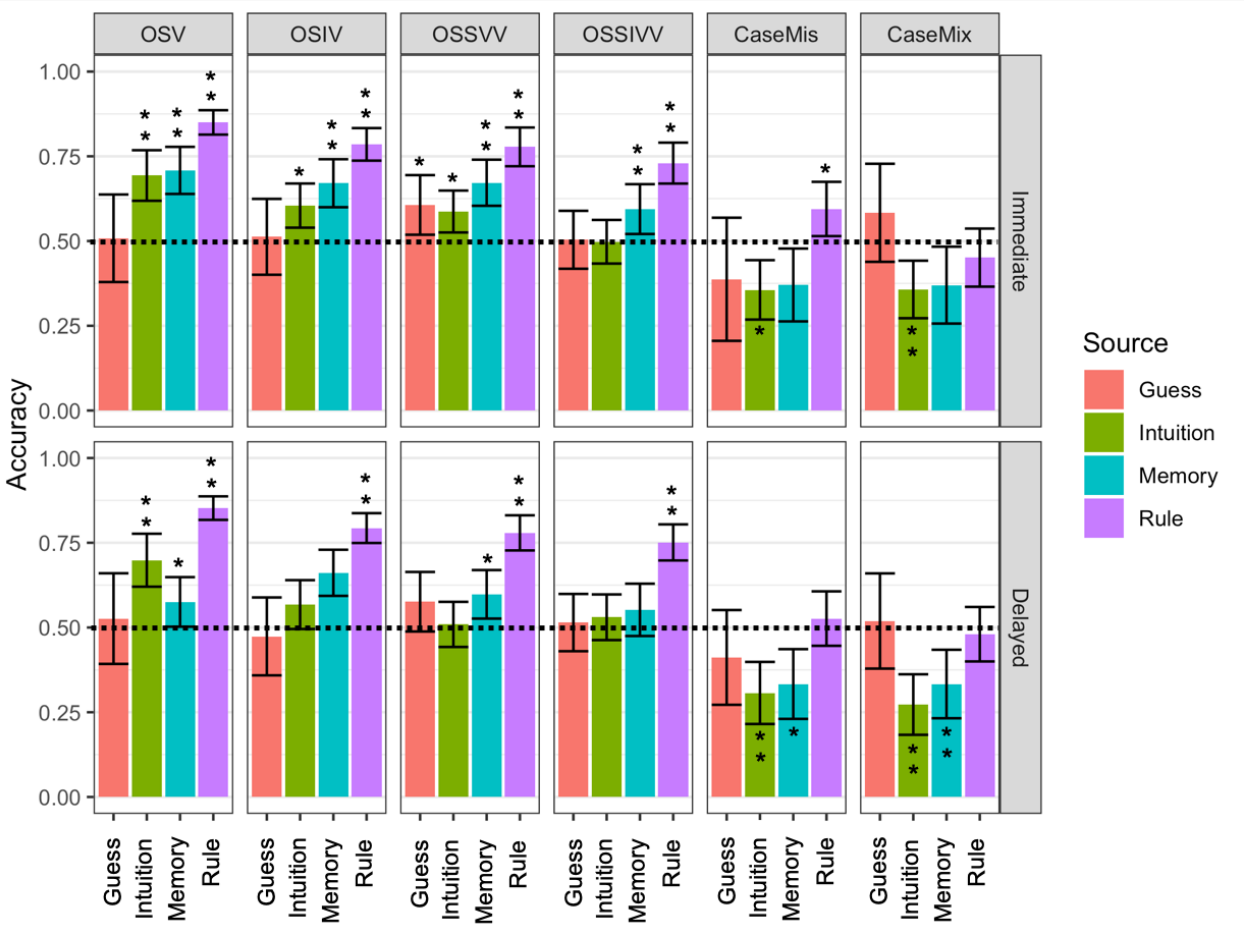


FIGURE 3. Results of source attributions at immediate and delayed posttest.
*Note*. * $p < .05$. ** significant after Bonferroni correction. Immediate posttest = upper panels and delayed posttest = lower panels. Error bars are 95% CIs.

For the source attribution data, the overall picture presented in Figure 3 seems to show four patterns. First, Rule as a basis of knowledge was robustly correlated with accuracy of performance for most of the structures except the two case-marking types, for which the confidence was not strongly related to accuracy in Table 2. Second, Memory (of exposure items) was also related to the above-chance level performance but seemed to be constrained to the word orders only. Third, the participants performed above chance when they claimed to be drawing on their intuitions, but the results showed that this largely worked for the simple word order types only. Lastly, when the participants based their responses on a complete guess, they largely performed at or below chance, except OSSVV, for which they performed significantly above chance. See Table S17 in Appendix G for a statistical summary.

**Delayed posttest**. Results of the confidence ratings and the source attributions at the delayed posttest are summarized in Table 3 and Figure 3, respectively. The confidence level on correct responses was significantly higher than that on incorrect responses for all of the word order types except for OSIV, whose results did not reach significance after a Bonferroni correction was made. Again, the effect size of the role of awareness was small as the difference between the

confidence levels on correct and incorrect responses was quite small, $d = 0.35$ for OSV, $d = 0.25$ for OSSVV and $d = 0.23$ for OSSIVV. Lastly, there was no significant difference in the confidence levels for correct and incorrect responses on Case Missing and Case Mixing, and the participants performed significantly above chance for these constructions, $t(27) = 6.24$, $p < .001$, 95% CI [1.23, 2.44] for Case Missing; $t(27) = 5.80$, $p < .000$, 95% CI [1.00, 2.09] for Case Mixing.

Table 3. *Mean Confidence Levels on Delayed U-AGJT*

| Construction | Correct (*SD*) | Incorrect (*SD*) | *t* | *df* | *SE* | Cohen's *d* |
|---|---|---|---|---|---|---|
| OSV | 3.92 (1.17) | 3.48 (1.30) | 4.21** | 352 | 0.10 | 0.35 |
| OSIV | 3.57 (1.24) | 3.36 (1.29) | 2.16* | 480 | 0.09 | 0.16 |
| OSSVV | 3.30 (1.23) | 2.99 (1.16) | 3.57** | 633 | 0.08 | 0.25 |
| OSSIVV | 3.21 (1.28) | 2.92 (1.14) | 3.31** | 705 | 0.08 | 0.23 |
| CaseMis | 3.55 (1.25) | 3.34 (1.21) | 1.58 | 355 | 0.12 | 0.17 |
| CaseMix | 3.30 (1.31) | 3.38 (1.16) | -0.61 | 301 | 0.12 | 0.06 |

*Note*. * $p < .05$. ** significant after Bonferroni correction.

Three main findings stood out in the source attribution data. First, the participants were particularly accurate when they were drawing on rules they had formulated about the language. However, this was not as effective for Case Mixing, for which they performed as well with complete guessing as they did with the rules. Second, memory of exemplar items as a basis of knowledge only worked to some extent, especially for the simple word orders. Lastly, they performed above chance only on OSV when they were drawing on their intuitions about the language, but their accuracy was completely at chance or even statistically worse on the other structures as well as when they were merely guessing. See Table S18 in Appendix G for a statistical summary.

**Word Monitoring Task**

**Immediate posttest**. The reliability of the task based on the Spearman-Brown prophecy formula with a split-halves method was $\rho = .95$. Mean overall raw RTs of the experimental and the control group at the immediate posttest were 572.36 ms (*SD* = 176.30) and 611.38 ms (*SD* = 311.12), respectively.[6] For a descriptive summary of the mean RTs at the immediate posttest, see Table S19 and S20 in Appendix H. A MANOVA on the TransRTs showed that there was a significant main effect of Grammaticality at the multivariate level, $F(1, 94) = 3.03$, $p = .009$; Pillai's trace V = .16, but the main effect of Group and the interaction of Group and Grammaticality were not significant, $F(1, 94) = 0.49$, $p = .810$; Pillai's trace V = .03; $F(1, 94) = .122$, $p = .993$; Pillai's trace V = .03. This indicated that the experimental group and the control group did not differ in their word-monitoring latencies nor did they differ in terms of the grammaticality effect (i.e., detecting an ungrammatical element). Although the main effect of Grammaticality was significant, follow-up univariate ANOVAs did not show significant results for any of the structure types considered separately.[7] Figure 4 graphically summarizes the monitoring latencies of the groups.

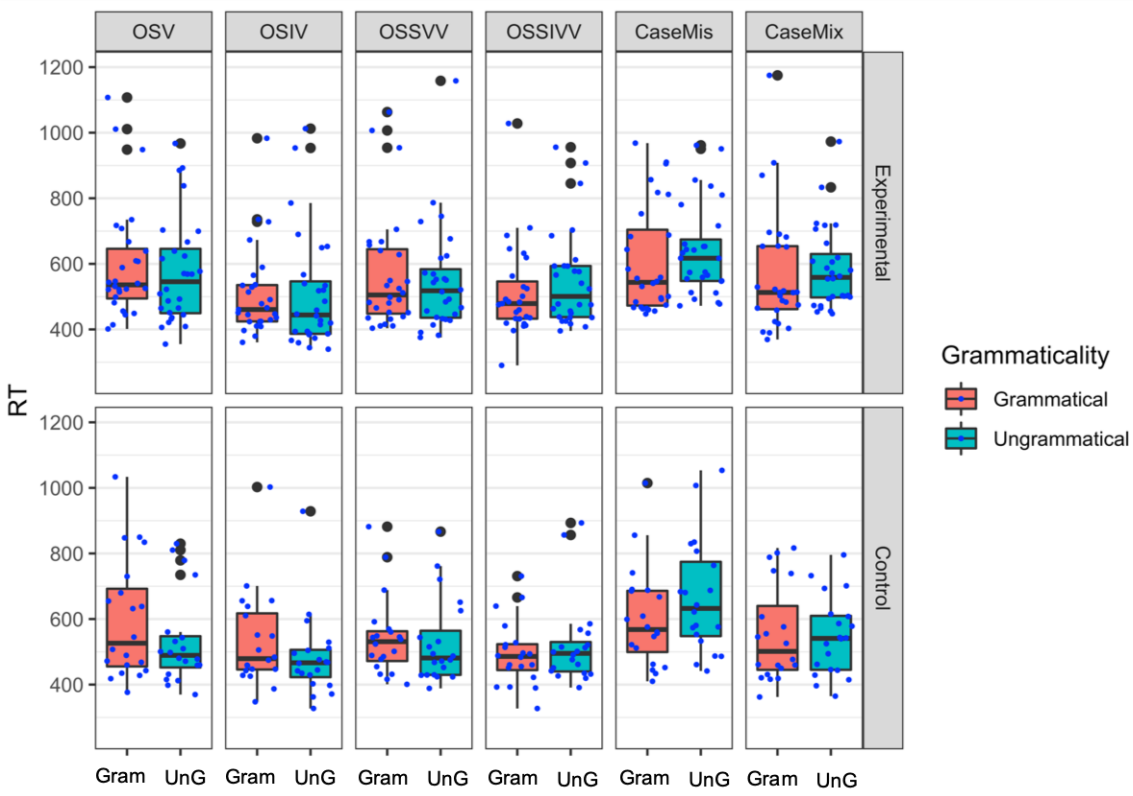


FIGURE 4. WMT latencies of experimental and control group at immediate posttest.
*Note*. Red = Grammatical and Blue = Ungrammatical

**Delayed posttest**. The reliability of the task based on the Spearman-Brown prophecy formula was $\rho$ = .81. Mean overall raw RTs of the experimental and the control group at the delayed posttest were 540.00 ms (*SD* = 170.64) and 541.21 ms (*SD* = 150.67), respectively (see Table S21 and S22 in Appendix H for a summary). Results of the MANOVA analysis showed that a main effect of Grammaticality was significant at the multivariate level, $F(1, 92) = 2.65$, $p = .020$; Pillai's trace V = .15, but a main effect of Group and the interaction of Group and Grammaticality were not significant, $F(1, 92) = 0.71$, $p = .641$; Pillai's trace V = .04; $F(1, 92) = 0.51$, $p = .795$; Pillai's trace V = .03, respectively. Follow-up univariate ANOVAs further showed that the main effect of Grammaticality was significant for the Case Missing items, $F(1, 92) = 6.02$, $p = .016$, $\omega^2 = .05$. This suggested that the two groups were comparable in their word-monitoring latencies as well as implicit knowledge assessed by sensitivity to grammatical violations, but this was the case only for the Case Missing items. Figure 5 graphically summarizes the monitoring latencies of the groups. Although follow-up repeated-measures ANOVAs were conducted for the experimental group and the control group separately to further examine the significant effect of Grammaticality on the Case Missing items, neither the main effect of Grammaticality nor its interaction with Construction were significant.

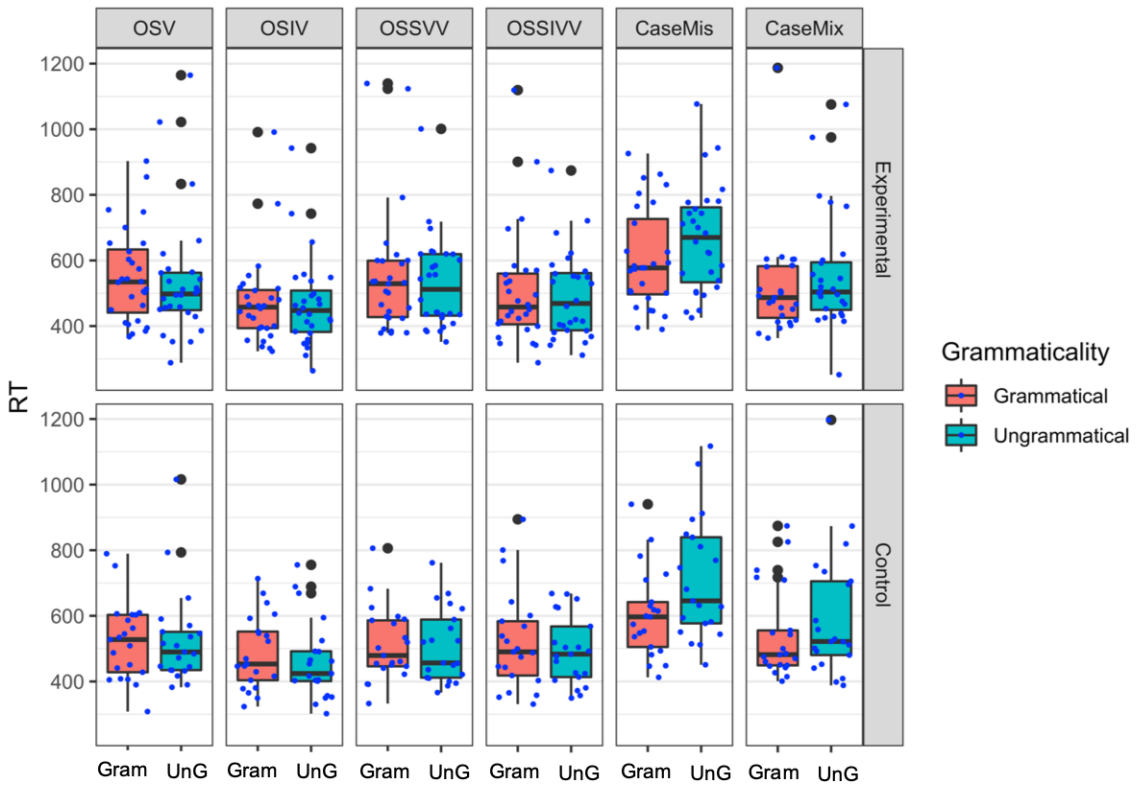


FIGURE 5. WMT latencies of experimental and control group at delayed posttest.
*Note*. Red = Grammatical and Blue = Ungrammatical

These results suggested that the participants, although feebly ($\omega^2$ = .05), showed implicit sensitivity to a grammatical violation, when one of the case markers was completely missing. The fact that not only the experimental group but also the control group showed the sensitivity seemed puzzling at first, but a careful reflection upon the exposure materials suggested that it was indeed logical. That is, although the control participants were exposed to the exemplar sentences whose word orders and positions of case markers were pseudo-randomized, it was also true that they never heard a sentence that was lacking a case marker. Hence, it can be expected that they learned the fact that the sentences carried three case markers, even without any understanding of their form-meaning relationships, and still showed the implicit sensitivity to the grammatical violation when the sentences were missing one of the case markers (but see Conclusion and Limitations for an alternative interpretation).

**Summary of Results**

The results of the experiment presented a quite complex picture, in which the triangulated methodological approaches differed in their sensitivity to development of EK and IK under incidental conditions, as summarized in Figure 6. It turned out that the objective measures were very stringent measures of L2 knowledge; in particular, WMT as a measure of IK. The results on WMT showed that the participants were implicitly sensitive to the case-missing violation in the

sentences, but they did not seem to possess any other IK. Subjective measures of awareness, however, suggested that the participants acquired both EK and IK of OSV, OSIV, OSSIVV, and CaseMis, but only EK for OSSVV and only IK for CaseMix.

| | Immediate | | | | | |
|---|---|---|---|---|---|---|
| | OSV | OSIV | OSSVV | OSSIVV | CaseMis | CaseMix |
| Objective | Explicit | Explicit | Explicit | Explicit | Explicit | Explicit |
| Subjective | Both | Both | Explicit | Both | Explicit | Implicit |
| | Delayed | | | | | |
| Objective | Explicit | Explicit | Explicit | Explicit | Both | Explicit |
| Subjective | Both | Explicit | Explicit | Explicit | Implicit | Implicit |

FIGURE 6. Summary of methodological approaches triangulated.

## Discussion

The current study triangulated subjective measures of awareness and objective measures of EK and IK to investigate whether adult learners develop EK and/or IK from brief incidental exposure to the target language. The results revealed clear discrepancies between the two approaches, in that while the subjective measures identified EK and IK of various Japlish constructions, the objective measures revealed that most of the knowledge was explicit, and that development of IK was minimal, only manifested in the detection of case-missing violations at the delayed posttest. The results from WMT are particularly interesting here; they are in a marked contrast with those of existing literature that claims to have found implicit learning of target structures (e.g., Chan & Leung, 2018; Grey et al., 2014; Rebuschat & Williams, 2012; Williams, 2005; but results are in fact consistent with Hama & Leow, 2010). Although evidence has accrued, it is crucial to note here that the criterion of learning in the previous studies was quite minimal, often operationalized as above-chance performance or an experimental group outperforming an untrained control group (but see Godfroid, 2016; Leung & Williams, 2011, 2012 for exceptions in employing RT-based measures to better tap into IK). Therefore, the question still remains to what extent such small learning effect approximates the kind of IK that L1 and advanced L2 speakers use during spontaneous communication. How robust is it? Can it be accessed automatically both in comprehension and production?

To reiterate, the current study *did* find evidence that adult L2 learners develop IK from brief incidental exposure. Case Missing, in particular, was an item type that was employed to measure knowledge of pure linguistic forms. It is thus concluded that implicit learning from brief exposure might be limited to learning of form knowledge or form-form associations only (see DeKeyser, 1995; Godfroid, 2016 for similar results). Of course, this is not to deny that implicit learning of form-meaning mappings may eventually be possible because such learning may happen after a much longer period of language exposure. It is known that implicit learning is the result of many thousands of language use experiences and practices (DeKeyser, 2003; N. Ellis, 1994; Robinson, 1997), and it thus comes as no surprise that implicit learning was quite limited in the present experiment (and perhaps in other studies) because it hardly afforded the amount of exposure that was sufficient for implicit learning of form-meaning mappings. What is needed is studies that analyze development of EK and IK in a more longitudinal time frame.

At the same time, however, it is crucial to understand why the two methodological approaches reached contradictory results. The present study hypothesizes that this is because the objective and subjective measures are able to detect IK at different stages of learning. Recall that a major difference between the two testing tasks was that the participants were afforded unlimited time to make their decision on the subjective measures, whereas their performance was required to be fluent and automatic on WMT. This leads to a possibility that they gauge IK that differs in degree of automaticity. This in fact explains why the previous validation studies of objective measures found evidence of IK (e.g., Suzuki, 2017; Vafaee et al, 2017), whereas very little of it was detected in the present experiment. While learners in Suzuki (2017), for instance, were highly

proficient advanced L2 speakers of Japanese with length of residence two years or longer, this study only afforded 200 exemplars of Japlish sentences. As research in cognitive psychology as well as SLA has attested that achieving automaticity is a long and slow process (e.g., Anderson, 1992; Logan, 1988; Perruchet & Pacteau, 1990) and that automatization of IK requires a larger number of instances of practicing the relevant structures (DeKeyser, 1995, 2003; N. Ellis, 1994; Robinson, 1997), it is no wonder that the participants were found to have developed almost no IK — because WMT required knowledge that must be available for automatic use; the subjective measures, on the contrary, tapped into IK that can be activated slowly (see Tamura et al., 2016 for unconscious but non-automatic knowledge), which the participants could develop from the brief training exposure.

The question one still needs to ask, however, is which of the two approaches provides us with an adequate measure of IK. Adopted from research in cognitive psychology, the subjective measures characterize IK as knowledge one is not aware of, defining it solely by the criterion of awareness (Dienes, 2008; Rebuschat, 2013). This implicates that they judge linguistic knowledge to be of an implicit kind if learners are deemed to be unaware of it, without consideration to its qualitative and functional aspects. In the field of SLA, however, IK has been conceptualized as the kind of knowledge that all native speakers possess (e.g., *acquired knowledge*: Krashen, 1982; *integrated knowledge*: Jiang, 2007) and it is the one that allows for fluent and spontaneous language use. Objective measures in fact capture this functional aspect of L2 knowledge because they are based on objective performance rather than subjective thoughts and introspections. It may well be, therefore, that IK, as has been defined in the field of SLA, can only be reliably gauged using objective measures such as WMT adopted in the present study. Subjective measures, on the contrary, carries a possibility of misclassifying EK as IK due to the sole reliance on the criterion of awareness, and even if it does so correctly, it can only detect IK that is at the initial stage of development. The criterion of (un)awareness, thus, might not be by itself sufficient to provide a full account of implicit L2 knowledge.

In further regards to IK, the results of the study replicated those of previous incidental learning studies in the sense that the effect of incidental exposure may not surface until some period after the exposure (e.g., Grey et al., 2014; Morgan-Short et al., 2012). This is indeed true in this study, which would have missed the implicit sensitivity to case-missing violations if it had not been for the delayed posttest. As claimed by Grey et al. (2014), adopting delayed posttests should be a methodological prerequisite for future studies. Yet, it is unclear as to why a delay is required for the effect of incidental exposure to emerge. Some researchers suggested that memory consolidation during sleep may play a role (e.g., Davis, Di Betta, Macdonald, & Gaskell, 2009; Walker, 2005) or it may be because the competition from explicit memory system diminishes over time (Poldrak & Packard, 2003).

Lastly, the results found that explicit learning was common under incidental conditions, and moreover, the direct comparison of the immediate and delayed posttest showed that the participants retained EK that they developed across the two testing sessions. This is clearly consistent with previous research on explicit L2 instruction which has repeatedly attested a

powerful potential of explicit learning mechanisms (e.g., Goo, Granena, Yilmaz, & Novella, 2015; Norris & Ortega, 2000); at the same time, however, these findings are also inconsistent with previous studies in that there were no differences between the experimental group's performance at the immediate and the delayed posttest, whereas the previous studies often found that EK declined after a delay with no training. One potential explanation is that it is because participants under incidental conditions (due to its inductive nature) may have to process L2 data more deeply than those who are given explicit rules before meaningful exposure (Hsieh et al., 2016; Leow, 2018). This implies that explicit knowledge does not necessarily diminish at the delayed posttest if leaners process L2 input deeply enough.[8]

## Conclusion and Limitations

The current study demonstrated that subjective measures of awareness and objective measures of EK and IK present conflicting evidence regarding whether adult L2 learners develop EK and/or IK from brief incidental exposure (see Figure 6). With regard to research design and methodology, the study also underscored the necessity to incorporate delayed posttests in research designs and to investigate development of EK and IK as a function of linguistic target and time (Grey et al., 2014). To this end, future investigations should also adopt new objective measures of IK (e.g., WMT, self-paced reading task, visual world paradigm) and seek to examine developing L2 knowledge from many different perspectives.

Before concluding, it is acknowledged that many conceptual and methodological shortcomings remained unaddressed in the study. First, the operationalization of awareness may have been too simplistic to capture its subtlety. There is an increasing recognition in the literature that awareness is graded rather than dichotomous in nature (see Leow, 2015 for a discussion). As Schmidt (1990) originally proposed, there can be different levels of awareness such that learners can be consciously aware of specific rules (i.e., noticing) but still remain unable to fully articulate them (i.e., understanding). In the present study, it is possible that the participants became aware of specific forms or rules, but its operationalization was not sensitive enough to detect it. In this light, what the subjective measures attributed to IK may in fact reflect the contribution of EK. The issue of subjective measures being insensitive to low level of awareness has been around for decades (Shanks & St. John, 1994), and the current study as well as many in the previous research are no exception.

Second, the use of a semi-artificial language was one of the limitations. The validity of semi-artificial languages has already been questioned by some researchers (see Godfroid, 2016; Leow, 2018 for instance): the cognitive processes for learning semi-artificial languages may differ from those of natural L2 learning due to increased saliency. The study adopted Japlish as it allowed for experimental control of the participants' experience with the language, but it was also likely that its artificial nature, in particular L1 words with unknown morphological markers attached to them could have triggered selective attention to such forms, eventually jeopardizing implicit processes that otherwise could have taken place. For example, the significant RT slow-down in

the delayed WMT can be alternatively interpreted as the participants consciously noticing the case-missing violation rather than as them being implicitly sensitive, because the state of an unknown case marker completely missing was so salient that they could consciously perceive it (exacerbated by the fact that they already knew the purpose of the study at the time of testing).

Lastly, the amount of exposure that the participants received was far less than ideal for them to develop IK. The fact that the participants were only able to develop IK of pure linguistic forms may be confounded by the small amount of exposure in the study, and it may not necessarily generalize to those who have used the target language for a longer period of time. In addition, such short-term training results in unnatural intensity of exposure to target constructions, which, in normal L2 learning, are often randomly distributed over a long period of time. To reiterate, the literature is in an acute need of research that conducts longitudinal analyses of developing EK and IK under incidental conditions. It is a well-known fact that SLA studies are typically short and tend to adopt testing materials that require conscious judgments; these study qualities are likely to bias findings about the effectiveness of explicit learning and instruction (e.g., Doughty, 2003; Norris & Ortega, 2000). Unless the literature accumulates empirical research that spells out longitudinal development of EK and IK, the true potential of implicit learning cannot be accurately evaluated.

**Notes**

1. An *incidental condition* here referred to an experimental paradigm in which participants are exposed to some domain of linguistic stimuli without being notified that the stimuli are governed by systematic rules or patterns, nor that there will be subsequent testing that will assess their learning of targets (Hulstijn, 2003). The study distinguishes incidental *learning* from incidental *condition*. Although some researchers use the terms interchangeably, it is important not to conflate the actual learning processes with learning conditions represented by linguistic environments surrounding the learner.

2. SPOT is a written fill-in blank test which explicitly asks learners to fill in blanks with target grammatical structures in written sentences.

3. Though previous research also focused upon other word order types, SOV, SIOV, SSOVV, and SSIOVV, a decision was made not to include them in the current investigation, so as to avoid their potential conflict with participants' L1 English word orders (i.e., SVO, SVIO, SVSVO, and SVSVIO).

4. Although the 5-point scale of the confidence ratings was adopted from Kachinske et al. (2015), the labels assigned to each option were our own. One anonymous reviewer rightly pointed out that the fact that there was no buffer between the top and lower score of each range (e.g., 3 = "confident: 70-80%", 4 = "very confident: 80-90%") could have been confusing to the participants. We agree. However, while we realize it is desirable to leave buffers when dividing scores into groups for further analysis, or for assigning participants to groups, etc., this situation is different in that the ranges are not outcomes, but ranges given to the participants to help them choose a category. Furthermore, we believe that the effect was quite minimal in our case because the results are highly consistent in that the participants were generally more confident on correctly answered items than incorrectly answered ones, except for Case-Mixing items, whose learning was smallest at the immediate ($\omega^2 = .08$) and delayed ($\omega^2 = .07$) posttest.

5. The same anonymous reviewer cautioned that the category of guessing in source attributions can be miscellaneous in nature, such that "guess" can comprise of both a random guess (i.e., 50/50%) and a conscious decision-making based on an educated guess. Although we tried our best to make our participants understood of each source attribution category, we acknowledge that there may have been some gap between what we expected participants to understand and what they actually did while performing the task. Against our operationalization of IK as reflected in guess responses, it may be that some of them were in fact based on an educated guess, which reflected EK.

6. The same reviewer pointed out that the fact that the experimental group generally responded faster than the control group can be attributed to the relatively high selection of rule and their use of explicit knowledge rather than implicit knowledge, as may have also been the case in Leung and Williams (2011, 2012). We would like to draw our readers' attention to such possibility, although examining its veridicality is out of scope for the present study.

7. Although it was feared that the non-significant results could have been caused by the small sample size of the study, 95% CIs of mean raw RTs showed that this was not the case (see Appendix H). Across grammatical and ungrammatical item types, they significantly overlap with each other for all construction types regardless of the group status or the timing of testing. We conclude therefore that the participants did not develop robust implicit knowledge (as measured by WMT) to begin with, and the non-significant results on WMT just reflect this fact.

8. We thank the same anonymous reviewer for suggesting this explanation. We must caution our readers, however, that the results should not be overgeneralized to L2 pedagogy. In reality, many rules of language are so complex that explicit induction (as was observed here) may not necessarily work in those situations.

**Supplementary Material**

Appendix A, B, C, D, E, F, G, H, and I
Results of Power Analyses

## References


Anderson, J. R. (1992). Automaticity and the ACT theory. *The American Journal of Psychology*, *105*, 166–180.

Chan, C. (1992). *Implicit cognitive processes: Theoretical issues and applications in computer systems design* (Unpublished Ph.D. thesis), University of Oxford, Oxford.

Chan, R. K. W., & Leung, J. H. C. (2018). Implicit knowledge of lexical stress rules: Evidence from the combined use of subjective and objective awareness measures. *Applied Psycholinguistics*, *39*, 37–66.

Cheesman, J., & Merikle, P. M. (1984). Priming with and without awareness. *Perception & Psychophysics*, *36*, 387–395.

Davis, M. H., Di Betta, A. M., Macdonald, M. J. E., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience*, *21*, 803–820.

DeKeyser, R. M. (1995). Learning second language grammar rules. *Studies in Second Language Acquisition*, *17*, 379–410.

DeKeyser, R. M. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313-348). Malden, MA: Wiley–Blackwell.

DeKeyser, R. M. (2009). Cognitive-psychological processes in second language learning. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 119–238).

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

DeKeyser, R. M. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 15–32). London: Routledge.

Dienes, Z. (2008). Subjective measures of unconscious knowledge. *Progress in Brain Research, 168*, 49–64.

Dienes, Z., Altmann, G., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1322–1338.

Doughty, C. J. (2003). Instructed SLA: Constraints, compensation, and enhancement. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 256–310). Oxford, UK: Wiley-Blackwell.

Ellis, N. C. (1994). *Implicit and explicit learning of languages*. San Diego, CA: Academic Press.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172.

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

Godfroid, A. (2016). The effects of implicit instruction on implicit and explicit knowledge

development. *Studies in Second Language Acquisition*, *38*, 177–215.
Godfroid, A., Kim, K. M., Hui, B., & Isbell, D. (2018, October). Validating implicit and explicit L2 knowledge measures: A research synthesis. Paper presented at the Second Language Research Forum, Montreal, QC.
Goo, J., Granena, G., Novella, M., & Yilmaz, Y. (2015). Implicit and explicit instruction in L2 learning: Norris and Ortega (2000) revisited and updated. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 443–482). Amsterdam: John Benjamins.
Grey, S., Williams, J. N., & Rebuschat, P. (2014). Incidental exposure and L3 learning of morphosyntax. *Studies in Second Language Acquisition*, *36*, 611–645.
Hama, M., & Leow, R. P. (2010). Learning without awareness revisited. *Studies in Second Language Acquisition*, *32*, 465–491.
Hamrick, P., & Sachs, R. (2018). Establishing evidence of learning in experiments employing artificial linguistic systems. *Studies in Second Language Acquisition*, *40*, 153–169.
Hsieh, H-C., Moreno, N., & Leow, R. P. (2016). Awareness, type of medium, and L2 development: Revisiting Hsieh (2008). In R. P. Leow, L. Cerezo, & M. Baralt (Eds.), *A psycholinguistic approach to technology and language learning* (pp. 131–150). Berlin: De Gruyter Mouton.
Hulstijn, J. H. (2003). Incidental and Intentional Learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Oxford, UK: Wiley-Blackwell.
Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language learning. *Language Learning*, *57*, 1–33.
Jiang, N. (2012). *Conducting reaction time research in second language studies*. New York: Routledge.
Kachinske, I., Osthus, P., Solovyeva, K., & Long, M. (2015). Implicit learning of a L2 morphosyntactic rule, and its relevance for language teaching. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 385–416). Amsterdam: John Benjamins.
Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759.
Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *20*, 79–91
Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Englewood Cliffs, NJ: Prentice Hall.
Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*, 294–340.
Leow, R. P. (2015). Implicit learning in SLA: Of processes and products. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages*. (pp. 47–67). Amsterdam: John Benjamins.
Leow, R. P. (2018). ISLA: How implicit or how explicit should it be? Theoretical, empirical, and pedagogical/curricular issues. *Language Teaching Research*, *22*, 1–18.

Leung, J. H., & Williams, J. N. (2011). The implicit learning of mappings between forms and contextually derived meanings. *Studies in Second Language Acquisition*, *33*, 33–55.

Leung, J. H., & Williams, J. N. (2012). Constraints on implicit learning of grammatical form-meaning connections. *Language Learning*, *62*, 634–662.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Morgan-Short, K., Finger, I., Grey, S., & Ullman, M. T. (2012). Second language processing shows increased native-like neural responses after months of no exposure. *Plos One*, *7*, e32974.

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

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or fragmentary knowledge? *Journal of Experimental Psychology: General*, *119*, 264–275.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 872–912.

Poldrack, R. A., & M. G. Packard. (2003). Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia*, *41*, 245–251.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*, 595–626.

Rebuschat, P. (2015). *Implicit and explicit learning of languages*. Amsterdam: John Benjamins.

Rebuschat, P., Hamrick, P., Riestenberg, K., Sachs, R., & Ziegler, N. (2015). Triangulating measures of awareness: A contribution to the debate on learning without awareness. *Studies in Second Language Acquisition*, *37*, 299–334.

Rebuschat, P., & Williams, J. N. (2012). Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics*, *33*, 829–856.

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

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 129–158.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*, 367–447.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*, 1229–1261.

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

Tamura, Y., & Harada, Y., Kato, D., Hara, K., & Kusanagi, K. (2016). Unconscious but slowly

activated grammatical knowledge of Japanese EFL learners: A case of tough movement. *Annual Review of English Language Education in Japan*, *27*, 169–184.

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.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*, 1–67.

Walker, M. P. (2005). A refined model of sleep and the time course of memory formation. *Behavioral and Brain Sciences*, *28*, 51–104.

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition*, *27*, 269–304.

Williams, J. N., & Kuribara, C. (2008). Comparing a nativist and emergentist approach to the initial stage of SLA: An investigation of Japanese scrambling. *Lingua*, *118*, 522–553.