Studies in Second Language Acquisition, 2015, **0**, 1–37.
doi:10.1017/S0272263115000455

VALIDATING GRAMMATICALITY JUDGMENT TESTS

Evidence from Two New Psycholinguistic Measures

QA Payman Vafaee, Yuichi Suzuki, and Ilina Kachisnke
University of Maryland

Several previous factor-analytic studies on the construct validity of grammaticality judgment tests (GJTs) concluded that untimed GJTs measure explicit knowledge (EK) and timed GJTs measure implicit knowledge (IK) (Bowles, 2011; R. Ellis, 2005; R. Ellis & Loewen, 2007). It has also been shown that, irrespective of the time condition chosen, GJTs' grammatical sentences tap into IK, whereas their ungrammatical ones invoke EK (Gutiérrez, 2013). The current study examined these conclusions by employing two more fine-grained measures of IK: that is, a self-paced reading task and a word-monitoring task. The results of a confirmatory factor analysis revealed that manipulating GJTs' time conditions and/or the grammaticality of the sentences does not render them distinct measures of EK and IK. The current work shows that GJTs are too coarse to be measures of IK, and that the different types of GJTs measure different levels of EK.

We would like to thank professors Steve Ross, Mike Long, and Robert DeKeyser and the *SSLA* editors and anonymous reviewers for their thoughtful suggestions and constructive feedback. We are also grateful to the participants of our study. All errors and omissions are, of course, our own.

Correspondence concerning this article should be addressed to Payman Vafaee, 9348 Cherry Hill Road, Apt #708, College Park, MD 20740. E-mail: payman.vafaee@gmail.com

INTRODUCTION

The current methodological study reexamined the second language (L2) knowledge type that nonnative English speakers draw on to perform grammaticality judgment tests (GJTs). Previous factor-analytic validity studies on GJTs employed elicited imitation (EI) and/or oral narrative (ON) tasks as measures of implicit knowledge (IK). Their authors concluded that manipulating GJTs' time conditions or the grammaticality of the sentences renders them distinct measures of IK and explicit knowledge (EK). Several studies yielded findings indicating that untimed GJTs measure EK, whereas timed GJTs measure IK (e.g., Bowles, 2011; R. Ellis, 2005; R. Ellis & Loewen, 2007). It has also been shown that, regardless of time condition, participant responses to different GJT stimulus types (i.e., grammatical and ungrammatical sentences) tap into IK and EK, respectively (Gutiérrez, 2013).

Unlike previous studies, the current study employed two new psycholinguistic measures of IK. More specifically, EI and ON were replaced by a word-monitoring task (WMT) and a self-paced reading task (SPRT), as these have been shown to be more valid measures of IK (Jiang, 2004, 2007; Jiang, Novokshanova, Masuda, & Wang, 2011; Suzuki, 2015; Suzuki & DeKeyser, in press). For this reason, it was hypothesized that, by including WMT and SPRT measures in a test battery, the results and conclusions pertaining to the construct validity of GJTs would be different from previous studies. In addition, a metalinguistic knowledge test (MKT), a well-established measure of EK, was included in the current study.

Through the comparison of the performance of learners on the WMT, SPRT, and MKT measures, as well as on several types of GJTs, it was possible to ascertain whether manipulating time conditions and/or grammaticality in GJT sentences can transform GJTs into distinct measures of IK and EK.

To foreshadow the results and conclusions yielded by this research, the aforementioned comparisons, carried out through confirmatory factor analysis (CFA), revealed that GJTs are too coarse to be measures of IK, and that manipulating their time conditions and sentence grammaticality does not render them distinct measures of IK and EK. Rather, we concluded that, as behavioral measures such as GJTs are not pure measures of IK or EK, on a continuum from being more explicit to more implicit, GJTs fall closer to the explicit end.

The following section situates the significance of the current study within a broader context of SLA research. It explains why more rigorous validation studies on GJTs are needed. The next section provides a critical review of previous validation studies on GJTs, as they motivate the research questions and design of the current study. AQ1

Explicit and Implicit Knowledge and the Interface Issue

The constructs of IK and EK are central to SLA theory construction and evaluation. There is a consensus that SLA draws on both implicit and explicit learning mechanisms (Bley-Vroman, 1991; DeKeyser, 2003; N. C. Ellis, 2005; R. Ellis, 2004), which in turn may result in IK and/or EK (Williams, 2009). Explicit knowledge is knowledge we are consciously aware of, whereas IK is the knowledge that we have but are not aware of (DeKeyser, 2009; Hulstijn, 2005; Williams, 2009). These two kinds of knowledge also differ in the extent to which one can or cannot verbalize them (R. Ellis, 2004, 2005). Our conscious access to EK allows us to verbalize it;[1] however, because IK is beyond awareness, it cannot be verbalized (DeKeyser, 2009).

It is currently believed that IK underlies the ability to use a L2 fluently and confidently; therefore, development of IK should be the ultimate goal of SLA (Doughty, 2003; N. C. Ellis, 1993; R. Ellis, 2005; Hulstijn, AQ2 2001).[2] Although there is solid evidence showing that implicit and explicit representations are neurologically distinct (Paradis, 2009; Ullman, 2011), the interaction between the two and how they influence each other are still subjects of controversy. In SLA research, this controversy is referred to as the *interface issue*. Central to the interface issue is to what extent explicit learning and instruction impact implicit learning and the development of IK (N. C. Ellis, 2011). There are three positions with regard to the interface issue—the noninterface, strong-interface, and weak-interface positions.

The noninterface position is often associated with Krashen (e.g., 1994), who contended that conscious learning about L2s and subconscious acquisition of L2s are two completely different phenomena that result in distinct sources of knowledge (EK and IK, respectively) with no AQ3 interface between them. According to this position, subconscious acquisition dominates L2 performance, learning can never convert into acquisition, and conscious learning can only be used as a monitor (editor) for performance. Proponents of the noninterface position believe that EK and IK are located in different areas of the brain and are thus accessed by different processes (Paradis, 1994).[3] AQ4

The strong-interface position, which is usually associated with DeKeyser (e.g., 2007), offers an opposite view. His strong-interface position AQ5 should be understood within the models of skill acquisition, such as ACT-R (Anderson & Lebiere, 1998). Within these models, a distinction AQ6 is made between declarative and procedural knowledge. According to DeKeyser (2009), “Declarative knowledge is knowledge THAT something is,” and procedural knowledge is “knowledge HOW to do something” (p. 121). According to skill-acquisition models, learners first develop a declarative encoding, whereby extensive practice is required to ensure that declarative knowledge leads to procedural knowledge. Further practice leads to automatized knowledge, which may not require any conscious processing (DeKeyser & Criado, 2013). This strong-interface position implies a causal relationship between declarative knowledge and proceduralization and automatization of the knowledge. In other words, EK forms a prerequisite for the generation of IK (Segalowitz & Hulstijn, 2005).

The main claim of the weak-interface position is that EK does not have a causal relationship with IK and only triggers or speeds up the implicit learning process, which subsequently leads to the generation of IK. For example, N. C. Ellis (e.g., 2005, 2008) contended that EK contributes indirectly to the acquisition of IK by promoting some implicit learning processes. According to the author, EK can make relevant linguistic features salient, thus enabling learners to notice them and recognize the gap between the input and the linguistic knowledge they possess (N. C. Ellis, 1994). N. C. Ellis (2008) further suggested that explicit and implicit processes work in tandem and that there is a dynamic interaction between them for consolidating IK.

The interface issue has been debated for decades and remains important to SLA research for theoretical and pedagogical reasons (Hulstijn, 2005). However, dealing with EK and IK constructs and testing rival interface hypotheses is challenging for several reasons. One of these challenges is the lack of reliable and valid measurement tools, which is particularly significant when attempting to measure IK. As both IK and EK sources are involved in L2 performance, it is almost impossible to devise pure behavioral measures of these two constructs (DeKeyser, 2009; R. Ellis, 2004, 2005). In addition, in constructing measures of IK, the operationalization of the concept of awareness and other challenges may be encountered. Consequently, researchers have been using a variety of imperfect measures (e.g., metalinguistic tests for EK and timed GJTs for IK). For this reason, rigorous validation studies on these measures are required.

Thus far, GJTs have been among the most extensively used measures in research on L2 acquisition. They have been subjected to several validation studies. However, due to several methodological limitations affecting previous studies, the construct validity of GJTs remains open to scrutiny. The following section highlights the limitations of previous studies and explains the motivation for the current project.

Extant Validity Research on GJTs

Previously, it was thought that GJTs provide a direct window into the learners' linguistic competence. However, it is currently acknowledged that GJTs only provide a measure of linguistic performance (R. Ellis, 1991). Many researchers have attempted to establish the type of linguistic knowledge—whether implicit, explicit, or a combination of both—learners draw on in their performance on GJTs. Earlier studies of this type yielded the conclusion that the nature of learners' knowledge, whether explicit or implicit, affects their judgment on GJTs (Chaudron, 1983). In other words, GJTs potentially lead participants to draw on both IK and EK, depending on what source of knowledge they *mostly* have at their disposal (R. Ellis, 2005).

Further research on GJTs aimed to investigate whether manipulating GJT designs leads to differential participant performance. For instance, it was hypothesized that if GJTs only ask participants to discriminate between well-formed and deviant sentences, it is possible that they evoke the use of pure intuition. On the other hand, the use of EK is very likely if GJTs require locating the error and even editing/correcting or describing the rule for judgment (R. Ellis, 1991). In addition, it was hypothesized that the kind of knowledge GJTs prompt learners to rely on to make their judgment depends on time conditions (i.e., whether the test is timed or untimed). Time pressure may encourage test takers to respond on the basis of their IK, whereas unlimited time may allow them to rely on their EK (Bialystok, 1979). Moreover, it has been hypothesized that the grammaticality of GJT sentences may induce differential performance because learners rely on IK and EK for judging grammatical and ungrammatical sentences, respectively (Gutiérrez, 2013).

Several factor-analytic validity investigations tested these hypotheses. This series of factor-analytic studies commenced with the psychometric work of Rod Ellis. R. Ellis (2005) conducted a psychometric study to design several measures of EK and IK and evaluated their construct validity with respect to EK and IK constructs. For operationalizing the two constructs and distinguishing between them, Ellis proposed seven criteria: degree of awareness, time available, focus of attention, systematicity and certainty, metalanguage, and learnability. He subsequently created five measures: namely, a timed GJT, an untimed GJT, an ON task, an EI task, and a MKT. Based on the seven criteria, Ellis predicted that the ON task, the EI task, and the timed GJT would tap into IK, whereas the untimed GJT and the MKT would evoke the use of EK.

R. Ellis (2005) submitted his test battery data to exploratory factor analysis (EFA). The results yielded a two-factor structure that confirmed his predictions. More specifically, the ON task, the EI task, and the timed GJT loaded on the first factor, whereas the untimed GJT and the MKT loaded on the second. Ellis labeled the two factors IK and EK, respectively. However, as Isemonger (2007) explained, Ellis's (2005) approach suffered from a few major flaws from a methodological and analytical perspective. As a result, the inferences and interpretations drawn were not well supported. For example, from a methodological and analytical point of view, because Ellis approached the factor analysis with an a priori hypothesis, the use of EFA was not acceptable. Instead, CFA should have been used because the prior hypothesis implied that the measures would measure the distinct constructs of EK and IK.

To overcome the analytical flaws of the approach employed by R. Ellis (2005), R. Ellis and Loewen (2007) reanalyzed the data used in that study through a CFA. In the CFA, they tested the two-factor model from the original EFA against a decision and production model, as a rival model. This is also a two-factor model, with EI and ON tasks loading on the production, and the two GJTs and the MKT loading on the decision factor. The results yielded by this approach showed that only the first hypothesized model produced adequate fit. However, whereas the CFA approach is indeed a better option than EFA, the authors' execution of the CFA could have been more thorough.

Confirmatory factor analysis is regarded as a process involving several stages—namely, (a) initial model specification, (b) parameter identification and estimation, (c) data-model fit assessment, (d) possible model modification, and (e) rival model identification (which may jeopardize causal inferences made from the original hypothesized model). However, the research conducted by R. Ellis and Loewen (2007) lacked a final CFA process—an adequate rival model identification. In addition, either alternative models were not specified in the GJT validation studies conducted to date or the tested alternative models were irrelevant to the main purpose of the study. When rival CFA models are not specified, conclusions about specific sets of measures can be highly compromised. As Isemonger (2007) explained, "It is important that rival models are tested because the fit of a particular model does not preclude the possibility that other untested models fit better" (p. 109).

The only rival model R. Ellis and Loewen (2007) tested was the decision and production model. The constructs in this rival model were, however, irrelevant to the main constructs of the study. One important rival model that could have provided a rebuttal to R. Ellis and Loewen's claims is a one-factor model, which enables directly testing whether the measures can actually distinguish the two constructs. Had this model been tested and had it fit the data as well as their two-factor model, it would not be possible for the researchers to conclude that their measures tapped into two distinct constructs of IK and EK.

More recently, Kachinske and Vafaee (2014) set out to examine a one-factor model as the alternative model to R. Ellis and Loewen's (2007) original model. They reanalyzed R. Ellis and Loewen's data and found that the one-factor model, which also accounted for the method effect (by correlating the error terms between similar tasks), fit the data as well as the authors' original model. As this finding suggests that the one-factor model is statistically as acceptable as R. Ellis and Loewen's two-factor model, the construct validity of the latter could not be supported. Kachinske and Vafaee further pointed out similar flaws in the factor-analytic approach adopted in several subsequent GJT validation studies (i.e., Bowles, 2011; Gutiérrez, 2013).

Bowles (2011) created a battery of five tests of Spanish as a second language by closely following R. Ellis's (2005) guidelines. Among other analyses, Bowles conducted a CFA to examine the factorial structure of her test battery and reported results that concurred with those obtained by R. Ellis and Loewen (2007). However, Bowles did not examine any rival CFA models against her two-factor model. Furthermore, the fitted two-factor model indicated high correlation between the two hypothesized factors ($r = .87$). With such a high correlation between the factors, it is hard to claim that the two factors are distinct. This reemphasizes the importance of testing rival models. If a single-factor model had been tested, results could have shown a good model fit, and the interpretations of the study could have changed.

Gutiérrez (2013) employed a similar test battery for a different L2 population—that is, Spanish learners in the United States. Gutiérrez followed the guidelines proposed by R. Ellis (2005) and created a battery of tests consisting of a timed GJT, an untimed GJT, and an MKT. The only test in the battery hypothesized to be a measure of IK was the timed GJT, whereas the untimed GJT and the MKT were considered measures of EK. The author simultaneously examined the role of time pressure on GJTs (timed and untimed) and the types of test items (grammatical and ungrammatical) in order to scrutinize GJTs as measures of IK and EK. It was hypothesized that, irrespective of time conditions, judging the grammatical sentences in GJTs taps into IK, and judging ungrammatical sentences engages EK. Gutiérrez conducted both an EFA and a CFA to test his hypotheses.

In the CFA, Gutiérrez (2013) tested two rival models. In the first, both the grammatical and ungrammatical sentences of the timed GJT loaded on the construct of IK, and the grammatical and ungrammatical sentences of the untimed GJT and the MKT loaded on the construct of EK. In the second model, the grammatical sentences of both the timed and untimed GJTs loaded on the construct of IK, and the ungrammatical sentences of both types of GJTs and the MKT loaded on the EK construct. The analyses yielded a better fit for the second model, implying that, regardless of time pressure, grammaticality of the stimulus is what distinguishes between the use of IK and EK in performing GJTs. However, Gutiérrez's study was not free from limitations.

First, having different types of GJTs as the only measures of IK, and a MKT as the sole measure of EK, limited Gutiérrez's ability to test important rival models. Second, bivariate correlations revealed statistically significant coefficients between the MKT and all types of GJTs, irrespective of stimulus type and time condition. Although the correlations between the MKT and the ungrammatical sentences of both types of GJTs were the greatest, the correlations between the MKT and the other measures cannot be ignored. Therefore, a one-factor model that accounted for method effects could have been a plausible rival model.

In summary, factor-analytic studies, including those conducted by Bowles (2011), Ellis (2005), Ellis and Loewen (2007), and Gutiérrez (2013), suffer from methodological issues, especially the failure to test rival models. This omission in the approach compromises any conclusions about the structural relations among measures and the factors in a model. Therefore, in future studies, several important rival models (e.g., one-factor models) should be tested.

Measuring IK through Online Processing

Psycholinguistic research has revealed how L2 learners process grammatical structures in real time (e.g., Clahsen & Felser, 2006; Kaan, 2014). Online grammatical processing has often been examined through reaction time (RT) measures, such as self-paced reading tasks (e.g., Jiang, 2004; Roberts & Liszka, 2013) and word-monitoring tasks (Granena, 2013; Jiang, Hu, Lukyanchenko, & Cao, 2010; Suzuki & DeKeyser, in press). These tasks can examine whether L2 learners are sensitive to grammatical violations while they are reading or listening for comprehension. In the self-paced reading task, for instance, participants read a sentence containing a target grammatical structure word by word, and the RT to each word read is recorded. Researchers examine whether participants slow down to read the word(s) once they encounter a grammatical error. For instance, when participants read a sentence with a subject-verb agreement violation like "The boy in the room enjoy reading many books," they will slow down when reading the word after the violation (i.e., *reading*), compared to their performance on the grammatical version of the same sentence (The boy in the room enjoys reading many books). By computing the RT difference between the grammatical and ungrammatical sentences, the sensitivity to grammatical violation can be estimated.

These online sentence-processing tasks are promising measures for tapping into IK (Suzuki & DeKeyser, in press). First, the tasks can capture sensitivity to grammatical violation as the sentence unfolds in real time. This minimizes the possibility that participants access their linguistic knowledge consciously, that is, that they rely on EK (Paradis, 2009). Second, they can direct attention to meaning, as each sentence is followed by a comprehension question asking about the content of the sentence. This second feature contrasts with the design of GJTs because GJTs of any kind draw participants' attention to form. When learners are instructed to decide whether a sentence is grammatical or ungrammatical, they inevitably pay attention to the form, which potentially invokes the use of EK. Although time pressure makes using EK harder, it does not rule out the possibility that EK is accessed. Therefore, the validity of GJTs can be further scrutinized if online processing measures that draw learners' attention to meaning are included in the current study.

The present investigations hold promise for scrutinizing the validity of GJTs because recent evidence suggests that online processing measures tap into IK. Suzuki and DeKeyser (in press) compared the word-monitoring task and the EI (a previously attested measure of IK) as measures of IK. The tasks employed in their study—along with a MKT as a measure of EK and a probabilistic serial reaction time (SRT) task, which served as a measure of aptitude for implicit learning—were administered to Japanese L2 learners who live in Japan. The study results showed that the word-monitoring task was related to the SRT task only, whereas the EI was associated with the MKT only. This pattern was found only among the learners who had lived in Japan longer than a certain number of years. This finding suggests that the word-monitoring task can serve as an implicit processing measure among learners with sufficient naturalistic L2 exposure.

The findings reported by Suzuki and DeKeyser also provide some implications for participant selection. Participants in the validation studies may need to have had sufficient naturalistic learning experience. For example, participants in the studies conducted by Gutiérrez (2013) and Zhang (2015) were classroom learners with limited exposure in L2 environments. As noted earlier, behavioral measures potentially prompt learners to draw on both IK and EK, depending on what source of knowledge they (mostly) have at their disposal (R. Ellis, 2005). Therefore, more rigorous validation studies on GJTs should at least recruit participants that are typically exposed to naturalistic, as well as classroom-based, learning opportunities.

THE STUDY

The current study investigated the construct validity of GJTs. By keeping modality constant, in the written mode, the study employed GJTs with combinations of different stimulus types and time conditions. Grammaticality judgment tests with two stimulus types (grammatical vs. ungrammatical sentences) and two time conditions (timed vs. untimed) were developed.

In order to advance the current understanding of the methodological problems in measuring EK and IK, the study incorporated two new psycholinguistic measures, along with the GJTs. They were a SPRT and a WMT, which should draw on IK to a greater extent. Explicit knowledge was operationalized as the use of linguistic knowledge with attention to form, requiring the use of metalinguistic knowledge. On the continuum of implicit to explicit linguistic processing, a MKT was employed as the most explicit test form.

RESEARCH QUESTIONS AND HYPOTHESES

The current study sought to address the following research questions:

> *Question 1.* What is the relationship among performance on different types of GJTs, a SPRT, a WMT, and a MKT?
> *Question 2.* Does manipulating the time condition and stimulus type in GJTs result in two distinct measures of EK and IK?

It was hypothesized that because GJTs draw attention to form, manipulating their time condition or stimulus type does not transform them into measures of implicit knowledge. Rather, online sentence-comprehension tasks—such as WMTs and SPRTs, which draw attention to meaning—are more valid measures of IK. It was further hypothesized that the ungrammatical sentences of both GJTs are more valid measures of knowledge of the target structures under examination. Therefore, it was posited that a CFA model—which includes (a) only ungrammatical sentences from both GJTs, as well as the MKT, as measures of explicit knowledge and (b) the WMT and SPRT as measures of implicit knowledge—would provide the best fit to the study data. Given the need to answer the research questions and test the hypotheses, CFA was chosen as the most suitable data analysis method.

METHOD

Participants

The main study participants were 79 learners of English as a second language, who started learning English after about the age of 10[4] in a formal setting and subsequently moved to the United States. These learners were Chinese international students who had lived in the United States for a minimum of 1 year. Chinese international students were chosen as study participants as this ensured that they all shared a common first language (i.e., this element was constant). These participants had scored a minimum of 90 on the TOEFL iBT test (or 6.5 on IELTS). According to the ETS website, 90 is the minimum score of the TOEFL iBT for the advanced level. In addition, according to ETS, an IELTS score of 6.5 is equivalent to the TOEFL iBT score of 90. In terms of the gender distribution of the sample, 52 of the participants were female, whereas the remaining 27 were male. Their educational attainment was mostly similar: 16 participants were undergraduates, whereas the remaining 63 were graduate students enrolled in various degree programs at a U.S. mid-Atlantic university. Table 1 presents descriptive statistics pertaining to students' demographic background.

In addition, the language tests employed in the study were piloted with a group of English native speakers (NSs), all of whom were undergraduate students, on two separate occasions. First, 20 NSs took part in the initial item analyses. In light of the results of the first phase, some of the materials and test items were modified, and the revised version was offered to 10 further NSs.

Target Structures

Four English target structures—present hypothetical conditional, third-person *s*, simple past/present perfect, and mass/count nouns—were used to construct the GJTs and the SPRT, WMT, and MKT measures. The reasons for choosing these four structures were twofold. First, past research (e.g., R. Ellis, 2009) suggested that these four target structures are among the most difficult structures in the English language for EFL and ESL learners to master. In addition, these structures could easily be incorporated into SPRT and WMT items.

Instruments

Before describing the details of each of the tasks, it should be mentioned that four parallel sets of sentences (Set 1, 1', 2, and 2') were created for each of the following tasks: the untimed GJT, the timed GJT, the

Table 1. Descriptive statistics for background information

	Mean	Median	*SD*	Min	Max	Range
Age	24.45	24	3.5	18	36	18
LOR[a]	31.85	26	25.4	12	146	134
AOA[b]	21.74	22	2.86	17	31	14
AOS[c]	9.45	10	2.96	1	18	17
TOEFL	98.46	99	5.43	90	110	20

[a] Length of residence in the United States in months.
[b] Age of arrival in the United States in years.
[c] Age of starting to learn English in China in years.

AQ10 SPRT, and the WMT. The sets were equal in terms of length (length of sentences was kept between 9 and 13 words) and complexity (all sentences were simple in structure, with no embedded clauses), as well as in the frequency and density of their lexicons (the frequency of the words was checked in a corpus).[5] The following samples targeting mass/count nouns illustrate how sentences in these four sets were constructed. In Set 1 and 1', the sentences were very similar, with minimal changes to some of the words:

1: Mary added a lot of sugar(s) to her coffee.[6]
1': Tom added a lot of sugar(s) to his tea.

The same level of similarity existed between Set 2 and 2`:

2: Mary likes to put a lot of sugar(s) to her coffee.
2': Tom likes to put a lot of sugar(s) in his tea.

However, when the differences between the sets were compared, the difference between Set 1 and Set 2 was found to be greater than that between 1' and 1 and 2 and 2', respectively. Thus, Sets 1 and 2 were used for the SPRT and WMT, whereas Sets 1' and 2' were employed in the untimed and timed GJTs. By using less similar sentences (e.g., Sets 1 and 2) for more similar tasks (e.g., the SPRT and WMT), spurious correlations between scores on the measures were avoided. In other words, relationships between the scores, if found, should indicate the commonality of the task designs (and constructs measured) rather than the similarity among the sentences used in each task. The tests were programmed and delivered through DMDX (Forster & Forster, 2003).

Timed and Untimed GJTs

Each of the timed and untimed GJTs were composed of 96 sentences. Sixteen items were presented for each target structure, half of which were grammatical and the other half ungrammatical. Similarly, among the 32 filler sentences (testing other target structures) included, 16 were grammatical and 16 ungrammatical. The results pertaining to the filler items were not included in the analyses. For each of the GJTs, two counterbalanced lists of sentences were created. In List 1, half of the target sentences were grammatical, and half ungrammatical. The grammaticality of the sentences was reversed in List 2, to ensure that no target sentence was presented twice in the same condition in one list.

For the untimed GJT, participants were instructed to decide whether the sentences were correct or incorrect and were reminded that there was no time constraint. Sentences in the timed GJT appeared on the screen for 2–5.5 s.[7] The time limit for each item for the main part of study was established based on the reaction time (RT) of the NSs in the pilot. More specifically, the time limit for each of the sentences of the GJTs was equal to 1.2 × NSs' RT.

SPRT

The SPRT assessed online grammatical sensitivity while participants were reading sentences for comprehension. In this task, participants were asked to read a sentence word by word as quickly and accurately as possible. The first word in a sentence appeared on the left-hand side of the screen, and when the keyboard button was pressed, the next word appeared to the right of the preceding word, which disappeared on 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 a comprehension question.

To develop the SPRT, two lists of stimulus sentences were constructed, each consisting of 64 target sentences (16 for each structure). The two lists were counterbalanced, whereby one half of the target sentences were grammatical and the other half ungrammatical in List 1. In List 2, the grammaticality of the sentences was reversed, so that no target sentence was presented twice in the same condition in one list. As before, 32 grammatical sentences were also included in each list as filler sentences. All the sentences were followed by a comprehension question that required a simple yes/no response. The ratio between yes and no responses was kept equal. Once again, RT on the filler sentences was excluded in the analyses.

As in the study conducted by Jiang (2007), the region of interest, where RTs were compared between grammatical and ungrammatical sentences, was set at three different word positions (see the underlined words in Table 2): at the critical word (i.e., where the error occurred in the ungrammatical sentences) and at the two words immediately following the critical word (to capture spillover effects). The word preceding the critical region was also used as a baseline in order to ascertain that the reading time of the word before the critical region did not differ between grammatical and ungrammatical sentences. If participants were sensitive to the grammatical error that preceded the critical region, their reading time would be delayed at (some of) these three positions. Table 2 shows some examples of each of the target structures.

Table 2. Sample sentences with critical regions

Target	Sample sentence	Critical word
Count/mass	Mary added a lot <u>of</u> **<u>sugar(s)</u>** <u>to her</u> coffee.	Sugar(s)
Third-person *-s*	The boy in the <u>room</u> **<u>enjoy(s)</u>** <u>reading many</u> books.	Enjoy(s)
Present perfect	Last spring <u>he</u> **<u>(has)</u>** <u>planted many</u> roses in the garden.	Has
Present hypothetical	If I lived in Miami, <u>I</u> **<u>can/could</u>** <u>have a</u> house near the beach.	Can/could

WMT

Similarly to the SPRT, the WMT also assessed online grammatical sensitivity during reading comprehension. Instead of self-paced reading, in this test participants were instructed to read a sentence presented automatically word by word on the screen. Their task was to respond to a target word or a monitoring word that appeared at one of the locations within the sentence. First, they were presented with a monitoring word in the center of the screen for 2 s. Next, each word in the sentence appeared on the screen for 1 s. The respondents were instructed to press the keyboard button as soon as they saw the target word in the sentence.

The monitoring word was always placed after the relevant target structure in the critical stimulus sentences. The difference in the RT to the target word between grammatical and ungrammatical sentences provided the index for online grammatical sensitivity. The monitoring word was located in the same position as the critical word in the SPRT, allowing the effects to be compared fairly between the WMT and the SPRT. Table 3 provides some examples of each of the target structures.

Table 3. Sample sentences with critical words and monitoring words

Target	Sample sentence	Monitoring word
Count/mass	Mary added a lot of **<u>sugar(s)</u>** <u>to</u> her coffee.	To
Third-person *-s*	The boy in the room **<u>enjoy(s)</u>** <u>reading</u> many books.	Reading
Present perfect	Last spring he **<u>(has)</u>** <u>planted</u> many roses in the garden.	Planted
Present hypothetical	If I lived in Miami, I **<u>can/could</u>** <u>have</u> a house near the beach.	Have

As in the SPRT, two lists of stimulus sentences were constructed for the WMT. Each list consisted of 64 target sentences (16 for each structure). The two lists were counterbalanced, whereby List 1 was composed of an equal number of grammatical and ungrammatical target sentences, and the grammaticality of the sentences was reversed in List 2. Once again, no target sentence was presented twice in the same condition in one list. Similarly, 32 grammatical sentences were also included in each list as filler sentences. All sentences were followed by a comprehension question requiring a yes/no response, with an equal ratio between the two. Reaction times for the filler sentences were excluded in the analyses.

MKT

The MKT was constructed with 20 items pertaining to the same target structures as in the GJTs, the SPRT, and the WMT. Five sentences were used for each target structure. All sentences in this task were ungrammatical and were similar to the ungrammatical sentences used in the GJTs. For each item, a sentence appeared on the screen. Participants were informed that the sentences were all ungrammatical, and their task was to state the reason for the ungrammaticality and then provide the correct form. There was no time constraint in this task. A rubric (Appendix A) was developed for rating the learners' performance on the MKT. The rubric detailed all of the possible acceptable and unacceptable responses. According to this rubric, partial credit could be assigned to each response, with 1 point for correct explanation and 1 point for correction. A total score of 2 was assigned for a response that included both a correct explanation and correction. Two researchers used the rubric and rated the responses independently. Their ratings were subjected to Rasch analysis, and the ability logit produced by Rasch was used as MKT data for further analyses.

Procedure

The five linguistic measures (timed GJT, untimed GJT, SPRT, WMT, and MKT) were administered, starting with more implicit tasks and progressing to more explicit tasks. Although other cognitive measures were administered in the study, the results are not reported here. The participants took these cognitive tests between the linguistic measures to delay exposure to the target structures and to minimize any practice effect. All the measures were administered in a 2-hr session.

However, the battery was divided into two 1-hr blocks with a 15-min break between the two. Learners were paid $25 for their participation. Table 4 shows the order and time for each of the measures.

ANALYSES

Pilot Study Involving NSs

All linguistic measures employed in the study were piloted with English NSs. As noted previously, the pilot consisted of two separate assessments. In the first phase, 20 NSs were asked to check the test items for ease of comprehension and correctness. Item and reliability analyses were subsequently conducted to diagnose the problematic items in the GJTs and MKT. The RTs of the NSs on the timed GJT items were also computed, to set the time limit for individual sentences in the task for the main study. The SPRT and WMT data were then analyzed to ensure that the tasks captured NSs' sensitivity to the incorrect target structures. The first phase of the pilot study identified some problematic items in the timed GJT, the untimed GJT, and the SPRT, prompting appropriate revisions. The revised tasks were subsequently administered to another group of the NSs (n = 10), who took part in the second phase of the pilot study. Because the WMT and MKT data from the first phase showed that the task functioned as expected, they were not given to the second group of NSs.

AQ11

GJTs. All the items from both GJTs were scored dichotomously, and the results were recorded as zero or one. Through item and reliability analyses, items with error rates higher than 25% were flagged for revision. These items were revised to ensure that grammatical and ungrammatical sentences clearly functioned in the expected manner. When the second phase of the pilot, with 10 NSs, was carried out, the results revealed that none of the items had an error rate higher than 25%.

Table 4. Order and time of the measures in each of the two blocks

Block 1		Block 2	
Task	Time	Task	Time
Consent form	5 min	Timed GJT	10 min
WMT	20 min	Cognitive measure 2	10 min
Cognitive measure 1	15 min	Untimed GJT	20 min
SPRT	20 min	MKT	15 min
		Background questionnaire	5 min

The NSs' RTs in the second round were used to set the time limit for the individual sentences for the learners. Following the work of R. Ellis (2005), the NSs' RT to each individual sentence was increased by 20% to set the time limit. Depending on the length and complexity of the sentences, the time limit varied across items. The average RT for the entire test was 3.39 s, and RTs ranged from 2 to 5.5 s.

SPRT. After the first phase of the pilot (n = 20), some items on the SPRT were revised. Based on the statistical results and the NSs' feedback, several sentences were revised to make them sound more natural or unambiguously grammatical or ungrammatical. The revised SPRT was then given to another group of NSs (n = 10), and their RTs to the word prior to the target structure (Region 0) and the average RT in the critical region (the target word and the following two words) for both grammatical and ungrammatical sentences were measured. Two assumptions were tested for the SPRT as a linguistic measure—namely, for the NSs (a) the average RT to Region 0 should be statistically the same across grammatical and ungrammatical sentences, and (b) the average RT to the critical region for the ungrammatical sentences should be statistically greater than that measured for the grammatical sentences. The results showed that (a) participants read the words prior to the target structure similarly in both grammatical and ungrammatical sentences, and that (b) NSs, who have IK of the target structures, slowed down when they came across the incorrect use of target structures. Through four separate sets of paired-samples t-tests, RTs to Region 0 and the critical region for the four target structures in the grammatical and ungrammatical sentences were compared. Table 5 summarizes the results from the second phase of the pilot, with 10 NSs.

The results revealed that there was no statistically significant difference in RTs to Region 0 across grammatical and ungrammatical sentences. On the other hand, for all four target structures, the RTs to the critical region were statistically greater in the ungrammatical sentences than in the grammatical sentences.

WMT. In order to establish whether NSs showed online sensitivity to the target structures, the RTs obtained by the group of NSs in the first phase of the pilot (n = 20) were analyzed. Through four separate sets of paired-samples t-tests, RTs to the target word across grammatical and ungrammatical items were compared, and the results are summarized in Table 6.

The results revealed that RTs to the target word in the ungrammatical sentences were statistically greater than RTs in the grammatical sentences. These results confirm that the WMT captured the online sensitivity.

Table 5. Results of the paired-samples *t*-tests and descriptive statistics for the SPRT (NS data, second phase of the pilot)

	Grammatical		Ungrammatical				
Outcome	*M*	*SD*	*M*	*SD*	*n*	*t*	*df*
Third-person Region 0	255.5	21.13	258.5	24.2	10	−.35	9
Mass/count nouns Region 0	257.2	14.49	256.9	9.59	10	.07	9
Past/perfect Region 0	276.6	16.26	293.4	32.09	10	−2.1	9
Present hypothetical Region 0	326.8	69.89	311.2	58.85	10	.47	9
Third-person critical region	668.9	54.1	999.6	312.29	10	−3.02*	9
Mass/count nouns critical region	614.2	63.2	1,337.3	285.82	10	−8.37*	9
Past/perfect critical region	600.9	73.52	990.3	227.6	10	−6.31*	9
Present hypothetical critical region	720.4	121.66	1,599.3	408.22	10	−6.31*	9

* $p < .05$.

Main Study on L2 Learners

After the two-stage pilot testing, the revised test battery was administered to the 80 L2 learners. Even though 80 learners took all the tasks, the data pertaining to one of the participants had to be excluded because the computer failed to record the required information for several tasks. Thus the following analyses were conducted on a sample size of 79.

AQ12

Item and Reliability Analysis. To begin with, item analyses were conducted on all the items in both the timed and untimed GJTs, and analyses were repeated for the grammatical and ungrammatical sentences separately. Henceforth, for brevity, the following acronyms will be used for different combinations of GJTs: total timed GJT (T-GJT), timed GJT

Table 6. Results of the *t*-tests and descriptive statistics for the WMT

	Grammatical		Ungrammatical				
Outcome	*M*	*SD*	*M*	*SD*	*n*	*t*	*df*
Third person	527.6	117.58	670.65	237.47	20	−3.05*	19
Mass/count nouns	508.75	97.26	625.1	255.44	20	−2.63*	19
Past/perfect	421.3	88.14	621.15	352.91	20	−2.55*	19
Present hypothetical	415.95	71.61	589.3	316.34	20	−2.62*	19

* $p < .05$.

grammatical sentences (T-GJT-G), timed GJT ungrammatical sentences (T-GJT-U), total untimed GJT (Un-GJT), untimed GJT grammatical sentences (Un-GJT-G), and untimed GJT ungrammatical sentences (Un-GJT-U). For each of the GJTs, items with a zero or negative item discrimination (ID) index (item-total correlation) were deleted to improve the reliability. Eight items were deleted from the timed GJT, and twelve from the untimed GJT. Next, reliability was estimated (Cronbach's alpha) for the aforementioned GTJ measures, the SPRT, and the WMT. The interrater reliability of the two raters for the MKT was also computed. Two independent ratings for the MKT were subjected to a Rasch analysis, and the ability logit for individual learners was computed to generate data for subsequent analyses. Table 7 summarizes the reliability estimates for all the tasks of the study. The reliability estimates for the GJTs and the MKT in the current study were within the acceptable range (above .70, as suggested by Nunnally, 1978), and they all were within the ranges reported in the previous research (Bowles, 2011; R. Ellis, 2005; Gutiérrez, 2013; Zhang, 2015). Cronbach's alpha for the SPRT was high, whereas the WMT's internal consistency was lower but close to .70.

Descriptive Statistics. Following item analysis and exclusion of items with an inappropriate ID index, descriptive statistics were computed for all measures. Using multiple sources, such as statistical tests (Kolmogorov–Smirnov and Shapiro–Wilk tests) and graphs, the univariate normality of the measures was assessed. Depending on the severity of the skewness, two types of transformation are usually conducted. A square root transformation is carried out for data that differ moderately from the normal distribution, whereas a log transformation is more appropriate for data exhibiting substantial deviation (Tabachnick & Fidell, 2001). The T-GJT, T-GJT-U, Un-GJT-U, SPRT, and WMT data did not require any transformation. A square root transformation was carried out for the remaining measures because they differed from the normal

Table 7. Reliability estimates

Task	Number of items	Reliability estimate
T-GJT	56	.75
T-GJT-G	26	.74
T-GJT-U	30	.78
Un-GJT	52	.83
Un-GJT-G	22	.74
Un-GJT-U	30	.88
WMT	32	.65
SPRT	32	.95
MKT	20	Interrater reliability = .91

distribution moderately. No log transformation was conducted. Table 8 summarizes descriptive statistics for all the measures. As can be seen, the univariate skewness and kurtosis values for all measures were within the acceptable range of +/– 1, with the exception of kurtosis for SPR. However, when a multivariate normality test was conducted, the assumption of multivariate normality was met (skewness: z score = .76, p value = .447; kurtosis: z score = 1.31, p value = .19; skewness and kurtosis: chi-squared = 2.3, p value = .32).

Correlational Analysis. Before conducting CFA, in order to explore the relationships among the measures of interest for the present study, a Pearson product moment correlation analysis was conducted. Table 9 summarizes the results, presenting only the statistically significant correlations, for clarity. For the complete table of correlations, see Appendix B.

As can be seen in Table 9, the WMT and SPRT correlated with each other, but not with any other measures. The T-GJT correlated only with the grammatical and ungrammatical sentences contained within. The Un-GJT also correlated with the grammatical and ungrammatical sentences it was composed of, as well as with the MKT. The T-GJT-G correlated with the Un-GJTG only, but the T-GJT-U correlated with the Un-GJT, Un-GJT-U, and MKT. Finally, the Un-GJT-U correlated not only with the Un-GJT and T-GJT-U but also with the MKT.

Confirmatory Factor Analysis. Confirmatory factor analysis was chosen as the main method of data analysis in order to test the prior hypotheses. Unlike in previous studies, 20 different CFA models were tested. These models included the model hypothesized in this study and models employed in previous studies (i.e. Bowles, 2011; R. Ellis & Loewen, 2007; Gutiérrez, 2013; Zhang, 2015), as well as several other rival models. Table 10 summarizes information pertaining to all tested models. It also

Table 8. Descriptive statistics

Task	Mean	*SD*	Min	Max	Skew	Kurt
T-GJT	27.25	6.52	12	43	–.08	–.2
T-GJT-G	23.27	.76	21.42	25	–.05	–.16
T-GJT-U	9.3	4.85	0	21	.4	–.35
Un-GJT	47.3	.94	45.34	50	.25	.54
Un-GJT-G	20.88	.715	19.26	22	–.42	–.55
Un-GJT-U	18.42	6.15	0	29	–.68	.08
MKT	1.86	.36	.93	2.59	.02	.44
SPRT	10.72	298.91	–1,274	749	–.82	3.78
WMT	–13.71	188.29	–486	448	.03	.22

Table 9. Correlational matrix for all the measures

Task	T-GJT	T-GJT-G	T-GJT-U	Un-GJT	Un-GJT-G	Un-GJT-U	MKT	SPRT	WMT
T-GJT	—	**.65****	**.76****						
T-GJT-G		—			**.29***				
T-GJT-U			—	**.33****		**.49****	**.22***		
Un-GJT				—	**.48****	**.85****	**.45****		
Un-GJT-G					—				
Un-GJT-U						—	**.47****		
MKT							—		
SPRT								—	**.24***
WMT									—

** Correlation is significant at the 0.01 level (two-tailed).
* Correlation is significant at the 0.05 level (two-tailed).

AQ37

Table 10. Summary of the tested CFA models

Model	Explicit knowledge (EK)	Implicit knowledge (IK)	Language knowledge (LK)
1a (R. Ellis & Loewen, 2007)	Un-GJT-U, MKT	T-GJT, WMT, SPRT	
1b. (1a with correlated errors)	Un-GJT-U, MKT *Un-GJT-U/MKT*	T-GJT, WMT, SPRT *WMT/SPRT*	
2a (Bowles, 2011)	Un-GJT, MKT	T-GJT, WMT, SPRT	
2b (2a with correlated errors)	Un-GJT, MKT *Un-GJT/MKT*	T-GJT, WMT, SPRT *WMT/SPRT*	
3a (Gutiérrez, 2013)	T-GJT-U, Un-GJT-U, MKT	T-GJT-G, Un-GJT-G, WMT, SPRT	
3b (3a with correlated errors)	T-GJT-U, Un-GJT-U, MKT *T-GJT-U/Un-GJT-U*	T-GJT-G, Un-GJT-G, WMT, SPRT *T-GJT-G/Un-GJT-G /WMT/SPRT*	
3c (Gutiérrez, 2013)	Un-GJT-G, Un-GJT-U, MKT	T-GJT-G, T-GJT-U, WMT, SPRT	
3d (3c with correlated errors)	Un-GJT-G, Un-GJT-U, MKT *Un-GJT-G/Un-GJT-U*	T-GJT-G, T-GJT-U, WMT, SPRT *T-GJT-G/T-GJT-U and WMT/SPRT*	
4a (our hypothesized model)	Un-GJT-U, T-GJT-U, MKT	WMT, SPRT	
4b (T-GJT-U cross-loading)	Un-GJT-U, T-GJT-U, MKT	T-GJT-U, WMT, SPRT	
4c (Un-GJT-U cross-loading)	Un-GJT-U, T-GJT-U, MKT	Un-GJT-U, WMT, SPRT	
5	T-GJT-G, T-GJT-U, Un-GJT-G, Un-GJT-U, MKT	WMT, SPRT	
6	T-GJT, Un-GJT, MKT	WMT, SPRT	
7 (one-factor)			T-GJT-U, Un-GJT-U, MKT, WMT, SPRT[a]
8a (one-factor with correlated errors)			T-GJT-U, Un-GJT-U, MKT, WMT, SPRT *T-GJT-U/Un-GJT-U and WMT/SPRT*

Continued

Table 10. continued

Model	Explicit knowledge (EK)	Implicit knowledge (IK)	Language knowledge (LK)
8b (one-factor with correlated errors)			T-GJT-U, Un-GJT-U, MKT, WMT, SPRT *T-GJT-U/Un-GJT-U, MKT/Un-GJT-U and WMT/SPRT*
8c (one-factor with correlated errors)			T-GJT-U, Un-GJT-U, MKT, WMT, SPRT *T-GJT-U/Un-GJT-U, MKT/Un-GJT-U, MKT/T-GJT-G and WMT/SPRT*
9a (one-factor)			T-GJT-G, T-GJT-U, Un-GJT-G, Un-GJT-U, MKT, WMT, SPRT
9b (one-factor with correlated errors)			T-GJT-G, T-GJT-U, Un-GJT-G, Un-GJT-U, MKT, WMT, SPRT *T-GJT-G/T-GJT-U, Un-GJT-G/Un-GJT-U and WMT/SPRT*
9c (one-factor with correlated errors)			T-GJT-G, T-GJT-U, Un-GJT-G, Un-GJT-U, MKT, WMT, SPRT *T-GJT-G/Un-GJT-G, T-GJT-U/Un-GJT-U and WMT/SPRT*

Note. Correlated errors are italicized.
[a] No correlated errors.

shows which measured variables loaded on which factors, as well as the relationship between error terms for models in which method effect was accounted for. LISREL Version 9.1 (Jöreskog & Sorbom, 2012) was used for running CFA.

The models labeled 1a to 3d tested the models adopted in previous studies (i.e., Bowles, 2011; R. Ellis & Loewen, 2007; Gutiérrez, 2013; Zhang, 2015). The details of these models are not explained here but are available in Table 10. The table provides all immediately necessary information about each of these models. Regarding models from previous studies, a version of each with correlated error terms for the measured variables was also tested. In CFA, the specification of correlated errors is made on the basis of source or method effects, which explains the additional indicator covariation that resulted from common assessment methods (Brown, 2006). In other words, in correlating errors, the aim was to account for the method effect and to improve the model fit indices. The models from previous studies were alternative models to the model hypothesized in the present study.

The next stage of the analysis involved testing Model 4a, developed as a part of this study. In this case, the T-GJT-U, Un-GJT-U, and MKT loaded on the EK factor, and the SPRT and WMT loaded on the IK factor. Next, rebuttals to Model 4a were examined by testing rival models (Models 4b–9c). In Models 4b and 4c, the T-GJT-U and Un-GJT-U cross-loaded on both factors, respectively. In Model 5, all combinations of GJTs loaded on the EK factor, and in Model 6, the total score from both the grammatical and ungrammatical sentences of both GJTs loaded on the EK factor. In Models 5 and 6, the SPRT and WMT loaded on the IK factor.

Models 7–9c allowed testing rival models to Model 4a. In these models, all the measured variables loaded on a single factor, labeled language knowledge (LK). As can be seen in Table 10, in order to account for the method effect, the error terms of various measured variables were also correlated. Several versions of Model 4, in which the error terms of various measures were correlated, were subsequently tested. However, as these models did not improve Model 4 significantly, the related findings are not reported here (see Appendix C for the covariance matrix used for the preceding analyses).

To evaluate and compare the plausibility of the CFA models, a profile of model fit indices recommended by Hu and Bentler (1999) and Mueller and Hancock (2008) was used and reported: (a) the chi squared (χ^2), with its degrees of freedom and *p*-value; (b) the standardized root mean square residual (SRMR); (c) the root mean square error of approximation (RMSEA); the comparative fit index (CFI); (d) the normal fit index (NFI); (e) the nonnormed fit index (NNFI); and (f) the goodness-of-fit AQ13 statistic (GFI). For a model to be deemed a good fit to the data, it had to meet the following criteria: the chi-squared should not be statistically significant at a .05 level; the SRMR and RMSEA should be lower than .08 and .06, respectively; and values greater than .95 for the CFI and NNFI and .90 for the NFI and GFI indicate a good model fit. Table 11 summarizes the model fit indices for all CFA models examined.

As can be seen in Table 11, in terms of model fit indices, only six of the models (4a, 4b, 4c, 6, 8a, and 8b) fit the data well. Fit indices for the remaining models were outside the acceptable range.

Model 4a was the model hypothesized in this study and was compared to the other models with acceptable fit indices. These models are nested models (with the exception of Model 6, which employs different measures from the rest), and, therefore, a formal chi-squared difference test was conducted by $\Delta\chi^2_{(df_1 - df_2)} = \chi^2_{df_1} - \chi^2_{df_2}$ and was distributed as a chi-squared distribution with $df = df_1 - df_2$ (Mueller & Hancock, 2008). Table 12 presents the results yielded by the chi-squared difference tests.

As can be seen in Table 12, the chi-squared difference test was not statistically significant for any of the comparisons. These results suggest that no model is statistically different from any other. It should be

AQ38

Table 11. Summary of the model fit indices for the tested CFA models

Index	CFI	NFI	NNFI	GFI	RMSEA	SRMR	Chi-squared
Criterion	≤ .95	≤ .90	≤ .95	≤ .90	≥ .06	≥ .08	Nonsignificant
Model 1a	0	−30.41	−116.1	.24	1.73	0.53	*χ^2 = 945.12 , *df* = 4
Model 1b	.5	.6	−1.48	.95	.25	.09	*χ^2 = 11.94 , *df* = 2
Model 2a	0	−37.61	−144.9	.2	1.89	2.07	*χ^2 = 1,132.75 , *df* = 4
Model 2b	.37	.51	−2.16	.94	.28	.11	*χ^2 = 14.24 , *df* = 2
Model 3a	.77	.65	.63	.92	.1	0.08	*χ^2 = 23.317, *df* = 13
Model 3b	.86	.75	.7	.95	.09	.07	χ^2 = 16.35, *df* = 10
Model 3c	.22	.27	−.26	.88	.18	.14	*χ^2 = 47.74 , *df* = 13
Model 3d	.56	.55	.07	.91	.16	.11	*χ^2 = 29.75 , *df* = 10
Model 4a	.99	.90	.96	.98	.05	.06	χ^2 = 4.66, *df* = 4
Model 4b	1	.98	1.17	.99	0	.03	χ^2 = 1.09, *df* = 3
Model 4c	1	.97	1.13	.99	0	.04	χ^2 = 1.46, *df* = 3
Model 5	.76	.64	.61	.76	.1	.09	*χ^2 = 23.864, *df* = 13
Model 6	1	.91	1.19	.99	0	.05	χ^2 = 2.53, *df* = 4
Model 7	.9	.82	.8	.96	.1	.09	χ^2 = 8.93, *df* = 5
Model 8a	.96	.91	.88	.98	.08	.06	χ^2 = 4.45, *df* = 3
Model 8b	.99	.95	.95	.99	.05	.05	χ^2 = 2.38, *df* = 2
Model 8c	.86	.86	−.45	.97	.27	.07	*χ^2 = 6.6, *df* = 1
Model 9a	.68	.57	.53	.92	.11	.1	*χ^2 = 28.1, *df* = 14
Model 9b	.75	.66	.51	.94	.11	.09	*χ^2 = 22.4, *df* = 11
Model 9c	.88	.75	.78	.95	.08	.08	χ^2 = 16.24, *df* = 11

* χ^2 is significant at the .05 level.

Table 12. Chi-squared difference formal test results

Model 4a: χ^2 = 4.66, *df* = 4	**Model 4b:** χ^2 = 1.09, *df* = 3	$\Delta\ \chi^2$ = 3.57, *df* = 1, *p* value = .06
	Model 4c: χ^2 = 1.46, *df* = 3	$\Delta\ \chi^2$ = 3.2, *df* = 1, *p* value = .07
	Model 8a: χ^2 = 4.45, *df* = 3	$\Delta\ \chi^2$ = .21, *df* = 1, *p* value = .65
	Model 8b: χ^2 = 2.38, *df* = 2	$\Delta\ \chi^2$ = 2.28, *df* = 2, *p* value = .32

noted that Model 6 could not be statistically compared to the other models because it utilizes different sets of measured variables.

In the next step, the factor loadings in Model 4a were compared to those in all the other models (except for Model 6) that provided a good fit to the data. The model with higher and significant factor loadings is considered to fit the data better. Table 13 summarizes the factor loadings for these six models.

The examination of the factor loadings revealed that only Model 4a had significant loadings for all the measured variables. In Models 4b and 4c, the T-GJT-U and Un-GJT-U did not load on the IK factor significantly. Therefore, Model 4a, a more parsimonious model, is superior to the others. In Model 8a and 8b, SPRT and WMT did not load on the LK factor significantly. In Model 8a, the loading of the MKT was not statistically significant either. Finally, in Model 6, the T-GJT, MKT, and WMT did not load on their corresponding factors significantly. The only issue with Model 4a that needs further explanation is that, at 1.34, the loading of the Un-GJT-U exceeded 1, indicating that the error variance for this measured variable was negative. However, according to Jöreskog (1999), standardized coefficients (loadings) that are greater than 1, especially if they are not significant and are smaller than 2.8, are not an issue. In addition, in Model 4a, the correlation between the two factors was not statistically significant. These results lead to the conclusion that Model 4a, which was the hypothesized

Table 13. Factor loadings for Models 4a, 4b, 4c, 8a, 8b, and 6

Model	T-GJT	T-GJT-U	Un-GJT	Un-GJT-U	MKT	SPRT	WMT
Model 4a		*.36		**1.34	*.33	*.42	*.58
Model 4b		*.53 for EK and –.33 for IK		**1.07	**.43	*.42	**.57
Model 4c		**.5		**1.1 for EK and .47 for IK	**.44	*.37	**.65
Model 8a		*.71		*2.28	.20	.05	.08
Model 8b		**.25		**3.14	**.91	.05	.07
Model 6	.08		*2.63		.17	*.76	.32

** Loading is significant at the .01 level.
* Loading is significant at the .05 level.

model of the study, provides the best fit to the data. Figure 1 illustrates the best fitting Model 4a.

DISCUSSION

Construct Validity of the GJTs

The current study aimed at investigating whether manipulating the time condition and/or stimulus type transforms GJTs into distinct measures of EK and IK. Unlike the previous validation studies (Bowles, 2011; R. Ellis, 2005; R. Ellis & Loewen, 2007; Zhang, 2015), in this work it was hypothesized that GJTs of any kind are too coarse to be measures AQ14 of IK. Grammaticality judgment tests draw attention to form, and applying time pressure does not necessarily prevent L2 learners from accessing EK (Suzuki & DeKeyser, in press). In addition, contrary to what Gutiérrez (2013) proposed, here it was hypothesized that the ungrammatical sentences of GJTs would provide a more valid measure of (explicit) L2 knowledge of the target structures.

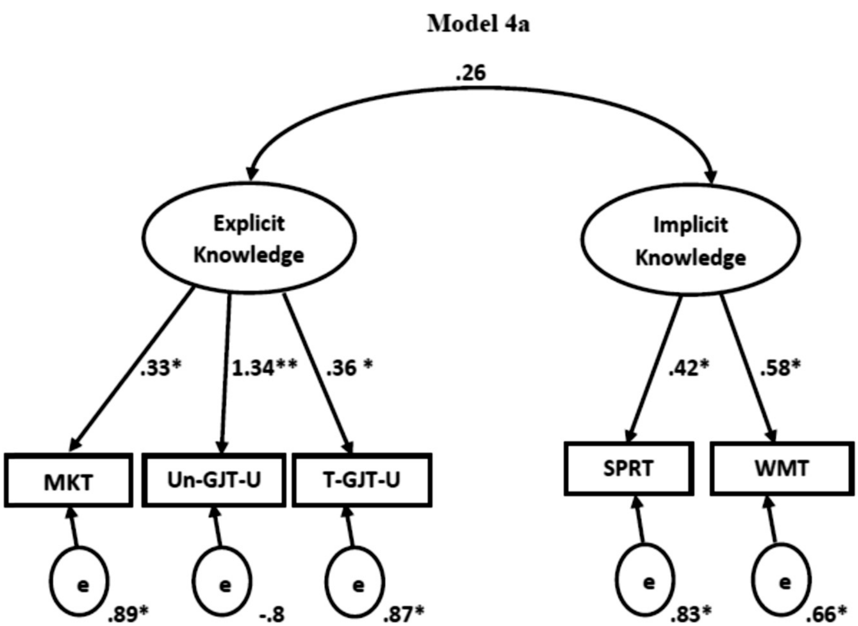

Figure 1. The best fitting CFA model (Model 4a). AQ39
Note. MKT = metalinguistic knowledge test, Un-GJT-U = ungrammatical sentences of the untimed GJT, T-GJT-U = ungrammatical sentences of the timed GJT, SPRT = self-paced reading task, and WMT = word-monitoring task.

To test these hypotheses, four different types of GJTs were developed: grammatical sentences in a timed GJT, ungrammatical sentences in a timed GJT, grammatical sentences in an untimed GJT, and ungrammatical sentences in an untimed GJT. In the tests, participants' total scores for the timed and untimed GJTs were calculated as well. Next, the learners' performance on the GJTs was compared with their performance on two other online processing measures (the WMT and SPRT), on one hand, and a well-established measure of EK (the MKT) on the other. Several CFA models were tested, including the ones adopted in previous studies, in which good fit indices were reported.

The CFA produced the best fitting two-factor model consisting of EK loaded onto the ungrammatical sentences of both the timed and untimed GJTs and the MKT and of IK loaded onto the WMT and SPRT. All models similar to those adopted in previous studies failed to achieve acceptable fit to the data, suggesting that the prior findings are of questionable validity. The bivariate correlation coefficients also showed that the grammatical sentences of both GJTs only correlated with each other, and not with any other measures. This finding may suggest that grammatical sentences in GJTs behave differently from the remaining measures employed in this work.

The correlation between the two factors in the best fitting model hypothesized in the current study (Model 4a) was small in magnitude and not statistically significant ($r = .26$). As noted by Brown (2006), "The size of the factor correlations in multifactorial CFA solutions should be interpreted with regard to the *discriminant validity* of the latent constructs. Small, or statistically non-significant, factor covariances are not usually problematic and are typically retained in the solution" (p. 131). The small, nonsignificant correlation between the factors in the current study can thus serve as evidence that the measures employed tapped into two distinct constructs. Previous studies, such as those conducted by Bowles (2011) and Zhang (2015), yielded large and significant correlations between the two factors in the models reported to have the best fit to the data. More specifically, Bowles (2011) reported a correlation of .87 between the two factors, whereas Zhang (2015) obtained .86. These high correlations weaken their argument that timed and untimed GJTs are distinct measures for EK and IK. This is consistent with Kachinske and Vafaee's (2014) conclusion that the data from the previous validation studies fit well in the one-factor model accounting for the method effect as well as in the two-factor models.

Given the small sample size in the current study, further post hoc analysis was conducted to assess whether the two factors in the best fitting model (Model 4a) can be recovered with the current sample size. MacCallum, Widaman, Zhang, and Hong (1999) and Velicer and Fava (1998) provide a thorough review of studies investigating the role of sample size in factor analysis. They demonstrated that the minimum sample size required largely depends on several factors, among which communalities were the most influential. According to MacCallum, Widaman, Preacher, and Hong (2001), "The level of communality has an especially strong interaction with N such that when communalities are high, good recovery of population factors can be achieved with relatively small samples" (p. 612). It has been suggested that low communalities in particular pose a serious problem in small sample sizes. The average communality was thus computed[8] for Model 4a, and it was .51. The value is above the low communality level (.35), as suggested by MacCallum et al. (2001), and this lends some support for the adequacy of this sample.

In sum, the present findings challenge the prior claims that timed and untimed GJTs are distinct measures of EK and IK constructs. Thus, based on the findings of the extant studies, it cannot be concluded that timed GJTs are measures of IK. If behavioral measures are considered to lie on a continuum from more explicit to more implicit, GJTs are probably considered closer to the explicit end of the continuum.

Using Online Processing Measures to Capture IK

The current study demonstrated that the online psycholinguistic measures (the WMT and SPRT) are tapping into a different construct than the one the GJTs are drawing on. This suggests that those newer tasks probably lie closer to the implicit end of the continuum relative to the GJTs. As delineated previously, online sentence-processing tasks can minimize the involvement of EK by capturing the online sensitivity to violations while attention is directed to meaning. Self-paced reading tasks and WMTs have been primarily utilized in psycholinguistic investigations (e.g., Clahsen & Felser, 2006) and have not been explicitly employed for addressing the issues of measurements of EK and IK until recently (Suzuki, 2015). The current findings corroborate those reported by Suzuki and DeKeyser (in press), who noted that real-time grammatical processing can index IK. In the work of Suzuki and DeKeyser (in press), the WMT was the only linguistic measure for IK and was not contrasted with the results of any other processing measure, such as a SPRT. The WMT adopted by Suzuki and DeKeyser (in press) was administered in the auditory modality,[9] whereas the WMT and the SPRT in the current study were given in the visual modality. Despite the modality difference, the critical design of the task was shared, and their performances converged. Combined, these results suggest that online comprehension tasks are good candidates for IK measures.

Recently, in SLA studies on implicit "learning," researchers have also started to apply RT-based online measures to assess to what extent implicit learning took place (e.g., Leung & Williams, 2011; Paciorek & Williams, 2015). Processing measures are more advantageous for capturing how L2 learners access linguistic knowledge in real time and can potentially reveal implicit learning processes. The recent applications of online measures in implicit learning research support the current finding that online comprehension measures draw on IK. Online processing measures offer great promise for further validation of EK and IK measures, as this study demonstrated that they can be employed to further scrutinize the validity evidence for the GJTs.

Suggestions for Further Research

The current study is not without limitations and offers several venues for further research. First, in order to investigate the construct validity of behavioral measures hypothesized to measure IK, the issue of awareness remains central. If the key definition of IK is that it definitively *does not involve* awareness, measures of awareness should be included in the validation studies. The current study did not employ any such measure of awareness.

Second, a test may be valid for certain purposes, such as a particular learning context or population of test takers, but not for others (Henning, 1987). The participants in the current study were learners with both classroom-based and naturalistic learning experience. The validity of GJTs may be assessed in different ways, depending on the specific context of the study in which GJTs are used. For instance, heritage learners with less formal instruction may perform on GJTs differently from the sample recruited for the present study, as they are posited to be less likely to possess EK compared to classroom learners.

Third, the current test battery only targeted four target structures. The original research by Rod Ellis targeted seventeen structures, with a similar number used in the subsequent studies. The smaller number of types of target structures tested in the current study may limit the generalizability of the present findings.

In addition, the current study focused only on Chinese L2 learners, whose L1 is typologically different from English. If the same test battery were administered to another L1 group, the results would likely be different. Because the explicit and implicit learning processes may be interactively influenced by the target structures to be acquired and learners' prior knowledge, including their L1 (e.g., DeKeyser, 2003; Leung & Williams, 2014; Williams, 2005), it may be worth expanding the current research to a different population and/or testing other linguistic structures.

Finally, in the model (Model 4a) that provided the best fit to the data, as well as in the remaining models, the factor loadings of the measured variables were relatively low. However, for stable latent variables to be measured, models are usually required to achieve higher factor loadings. Obviously, more rigorous further studies on the validity of GJTs are urgently needed.

CONCLUSION

The current study set out to investigate the validity of GJTs as measures of EK and IK. The data analyses were extremely stringent and included a thorough execution of CFA. Thus it was possible to reveal the limitations of the previous research. Specifically, the study provided evidence that challenged conclusions from previous studies that time pressure renders GJTs measures of IK. The claim that timed GJTs are measures of IK no longer holds when online comprehension tasks that can capture grammatical sensitivity are employed. Given the nature of GJTs—they involve focus on form—GJTs may be considered to be located closer to the explicit, rather than the implicit, end of the continuum (DeKeyser, 2003). Endorsing the recent call for the use of processing measures for capturing implicit processes (Andringa & Curcic, 2015; Suzuki & DeKeyser, in press), the current study demonstrated the potential of online comprehension tasks as measures of IK.

Received 31 March 2015
Accepted 10 November 2015
AQ15 *Final Version Received 3 September 2015*

NOTES

1. Although EK can be verbalized, obviously not everyone has the metalinguistic means to articulate the rules clearly and completely (DeKeyser, 2009). Therefore, lack of verbalization ability is not necessarily evidence that learners do not possess EK.

2. The idea presented previously is controversial, and several other SLA scholars (e.g., DeKeyser, 1997, 2009) believe that the ultimate goal of SLA—fluent and accurate use of a L2—can also be accomplished by using automatized EK. AQ40

3. In a strong, noninterface position, the possibility that EK can be transformed into IK and vice versa is completely ruled out (Bowles, 2011; Hulstijn, 2002). However, in a weaker version of this position, the possibility of the transformation of IK into EK is recognized (Bialystok, 1994). AQ41

4. The participants' responses to the question about the age at which they started learning English at school were variable. The minimum reported age was 1, and the maximum was 18. However, the average reported age was 9.75 years old, and the median was 10.

5. The corpus used in the current study was the Corpus of Contemporary American English (COCA): http://corpus.byu.edu/coca/. All words were taken from the 1,000 most frequent word families of English.

6. In some dialectics of English, this sentence may not be considered unambiguously ungrammatical. However, these sentences were piloted with NSs of American English, and all NS participants agreed that "a lot of sugars" is wrong.

7. Following previous studies of this type (e.g., Gutiérrez, 2013), this time limit was set to 3–6 s for NSs in the pilot.

8. The value was computed by the average of the squared factor loadings (Brown, 2006).

9. The target word was presented visually on the computer screen, but the carrier sentence was presented auditorily.

REFERENCES

Andringa, S., & Curcic, M. (2015). How explicit knowledge affects online L2 processing [Special issue]. *Studies in Second Language Acquisition*, *37*(2), 237–268. doi:10.1017/S0272263115000017

Bley-Vroman, R. (1991). The logical problem of foreign language learning. *Linguistic Analysis*, *20*(1–2), 3–49.

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

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

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*(1), 3–42.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum. AQ16

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. AQ17

DeKeyser, R. M. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 312–348). Oxford, UK: Blackwell.

DeKeyser, R. M. (2007). Skill acquisition theory. In B. V. J. Williams (Ed.), *Theories in second language acquisition* (pp. 97–114). Mahwah, NJ: Erlbaum. AQ18

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. 119–138). Oxford, UK: Wiley-Blackwell.

DeKeyser, R. (2012). Interactions between individual differences, treatments, and structures in SLA. *Language Learning*, *62*(s2), 189–200. AQ19

DeKeyser, R., & Criado, R. (2012). Automatization, skill acquisition, and practice in second language acquisition. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 4501–4504). Oxford, UK: Wiley-Blackwell. AQ20

Ellis, N. C. (1993). Rules and instances in foreign language learning: Interactions of explicit and implicit knowledge. *European Journal of Cognitive Psychology*, *5*(3), 289–318.

Ellis, N. C. (Ed.). (1994). *Implicit and explicit learning of languages*. Academic Pr. AQ21

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

Ellis, N. C. (2008). The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *The Modern Language Journal*, *92*(2), 232–249.

Ellis, N. C. (2011). Implicit and explicit SLA and their interface. *Implicit and Explicit Language Learning: Conditions, Processes, and Knowledge in SLA and Bilingualism*, 35–47.

Ellis, R. (1991). Grammaticality judgments and learner variability. In H. Burmeister & P. Rounds (Eds.), *Variability in second language acquisition: Proceedings of the tenth meeting of the Second Language Acquisition Forum* (pp. 25–60). Eugene: Department of Linguistics, University of Oregon.

Ellis, R. (1993). Second language acquisition and the structural syllabus. *TESOL Quarterly*, *27*, 91–113. AQ22

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

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

Ellis, R. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching* (Vol. 42). Bristol, UK: Multilingual Matters.

Ellis, R., & Loewen, S. (2007). Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger. *Studies in Second Language Acquisition*, *29*(1), 119–126. doi:10.1017/S0272263107070052

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

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, *63*(4), 665–703.

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition*, *35*(3), 423–449. doi:10.1017/S0272263113000041

Haig, J. (1991). *Universal grammar and second language acquisition: The influence of task type on late learner's access to the subjacency principle (*TESL monograph*)*. McGill University, Montreal, Quebec, Canada. AQ23

Han, Y., & Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research*, *2*(1), 1–23. AQ24

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Cambridge, MA: Newbury House.

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge, UK: Cambridge University Press.

Hulstijn, J. H. (2002). Towards a unified account of the representation, processing and acquisition of second language knowledge. *Second Language Research*, *18*(3), 193–223. 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*(2), 129–140.

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

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

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

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

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

Jöreskog, K. G. (1999). How large can a standardized coefficient be? Unpublished report. *SSI Central, Inc*. Retrieved from http://www.ssicentral.com/lisrel/techdocs/HowLargeCanaStandardizedCoefficientbe.pdf

Jöreskog, K. G., & Sorbom, D. (2012). LISREL 9.1 [Computer software]. Lincolnwood, IL: Scientific Software International.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, *4*(2), 257–282.

Kachinske, E., & Vafaee, P. (2014). *Reexamining the validity of GJTs as measures of implicit knowledge: Reanalysis of Ellis & Loewen (2007), Bowles (2011), and Gutierrez (2013)*. Paper presented at the Second Language Research Forum (SLRF), University of South Carolina, SC. AQ25

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342. AQ26

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*(3), 321–340. AQ27

Kline, P. (1994). *An easy guide to factor analysis*. London, UK: Rutledge. AQ28

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press. AQ29

Leung, J. H. C., & Williams, J. N. (2011). The implicit learning of mappings between forms and contextually derived meanings. *Studies in Second Language Acquisition*, *33*(1), 33–55.

Leung, J. H. C., & Williams, J. N. (2014). Crosslinguistic differences in implicit language learning. *Studies in Second Language Acquisition*, *36*(4), 733–755.

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. Philip, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 94–112). Tonawanda, NY: Multilingual Matters. AQ30

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, *36*(4), 611–637.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99.

Meara, P. M. (2005). *LLAMA language aptitude tests*. Swansea, UK: Lognostics. AQ31

Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. In J Osborne. (Ed.), *Best practices in quantitative methods* (pp. 488–508). Thousand Oaks, CA: Sage.

Murphy, V. A. (1997). The effect of modality on a grammaticality judgment task. *Second Language Research*, *13*(1), 34–65. AQ32

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Paciorek, A., & Williams, J. N. (2015). Implicit learning of semantic preferences of verbs [Special issue]. *Studies in Second Language Acquisition*, *37*(2), 359–382. doi:10.1017/S0272263115000108

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

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage. AQ33

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 585–594. AQ34

Roberts, L., & Liszka, S. A. (2013). Processing tense/aspect-agreement violations on-line in the second language: A self-paced reading study with French and German L2 learners of English. *Second Language Research*, *29*(4), 413–439. doi:10.1177/0267658313503171

Segalowitz, N., & Hulstijn, J. (2005). Automaticity in bilingualism and second language learning. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 371–388). New York: Oxford University Press.

Suzuki, Y. (2015). *Using new measures of implicit L2 knowledge to study the interface of explicit and implicit knowledge* (Unpublished doctoral dissertation). University of Maryland, College Park.

Suzuki, Y., & DeKeyser, R. M. (in press). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, *65*(4).

Tinsley, H. E., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology*, *34*(4), 414–424. AQ35

Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, *2*(10), 717–726.

Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*(2), 231.

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition*, *27*(2), 269–304. doi:10.1017/S0272263105050138

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

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

APPENDIX A

MKT RUBRIC

1. Third-person -*s*

   a. Full explanation should consist of = because the noun/subject/"the name of the word appearing in the sentence" is singular, you should use the verb + *s*/'*s* should be added to the verb/not plural.
   b. Participants will have to mention the singular noun or third-person singular and verb form.

2. Present perfect/simple past

   a. Full explanation should consist of = mention of simple past/an event occurring at a specific time in the past.

3. Countable/uncountable

   a. Full explanation should consist of = mention of specific terminology, such as countable/uncountable or can be plural/cannot be plural.

4. Hypothetical/second conditional

   a. Full explanation should consist of = the modal/the verb should be in the past tense to agree with the tense of the verb in the first sentence, or because the sentence describes a hypothetical situation/unreal past/hypothesis/assumption/supposition, or subjunctive mood.

APPENDIX B

AQ36

CORRELATIONAL MATRIX

Task	T-GJT	T-GJT-G	T-GJT-U	Un-GJT	Un-GJT-G	Un-GJT-U	MKT	SPRT	WMT
T-GJT	—	**.65****	**.76****	.22	.07	.22	.06	.02	−.13
T-GJT-G		—	.01	−.02	**.29***	−.21	−.14	.1	−.09
T-GJT-U			—	**.33****	−.16	**.49****	**.22***	−.07	−.11
UN-GJT				—	**.48****	**.85****	**.45****	.12	.16
UN-GJTG					—	−.01	.17	.02	.01
UN-GJT-U						—	**.47****	.13	.14
MKT							—	.08	−.03
SPRT								—	**.24***
WMT									—

APPENDIX C

COVARIANCE MATRIX

	T-GJT	T-GJT-G	T-GJT-U	Un-GJT	Un-GJT-G	Un-GJT-U	MKT	SPRT	WMT
T-GJT	42.448								
T-GJT-G	3.216	0.580							
T-GJT-U	23.896	0.025	23.496						
Un-GJT	1.343	–0.019	1.494	0.887					
Un-GJT-G	0.338	0.156	–0.550	0.321	0.511				
Un-GJT-U	8.611	–0.981	14.500	4.950	–0.051	37.810			
MKT	0.142	–0.038	0.385	0.152	0.044	1.028	**0.128**		
SPRT	42.648	22.157	–102.581	33.229	3.728	232.541	**8.417**	**89,347.691**	
WMT	–157.921	–13.305	–103.949	28.314	1.201	163.838	**–2.103**	**13,540.185**	**35,451.312**