Studies in Second Language Acquisition, 2013, **35**, 423–449.
doi:10.1017/S0272263113000041

THE CONSTRUCT VALIDITY OF GRAMMATICALITY JUDGMENT TESTS AS MEASURES OF IMPLICIT AND EXPLICIT KNOWLEDGE

Xavier Gutiérrez
University of Windsor

Grammaticality judgment tests (GJTs) have been, and continue to be, frequently used in the field of SLA as a measure of learners' linguistic ability in the second language (L2). However, only a few studies have examined their construct validity as measures of implicit and explicit knowledge (Bowles, 2011; R. Ellis, 2005), and even fewer have explored in detail how features of these tests, such as time pressure and task stimulus, affect their construct validity (Loewen, 2009). The purpose of this paper is to examine the effect that time pressure and task stimulus have on the type of knowledge representations on which L2 learners draw when performing GJTs. The results show that the grammatical and ungrammatical sections of a timed and an untimed GJT loaded differently in both exploratory and confirmatory factor analyses. This finding can be interpreted as indicating that grammatical and ungrammatical sentences constitute measures of implicit and explicit knowledge, respectively. Additionally, the results show that time pressure and task stimulus have significant effects on learners' performance on GJTs.

This research was funded by a Humanities and Social Sciences Research Grant from the University of Windsor. I would like to thank Rod Ellis, Shawn Loewen, and the anonymous *SSLA* reviewers for their thoughtful suggestions. I am also grateful to Jelena Primorac, Sandra Gidak, and Magdalene Friesen for their help in coding and analyzing the data, to the students who participated in the study, and to their instructors for allowing me access to their classes.

Correspondence concerning this article should be addressed to Xavier Gutiérrez, Department of Languages, Literatures, and Cultures, University of Windsor, 401 Sunset Avenue, N9B 3P4, Windsor, Ontario, Canada. E-mail: xavierg@uwindsor.ca

© Cambridge University Press 2013

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

The field of SLA has recently experienced a trend to attempt to ascertain the validity and reliability of instruments aimed at measuring the constructs of implicit and explicit knowledge (Bowles, 2011; R. Ellis, 2005; Ellis & Loewen, 2007). Although some instruments such as elicited imitation tasks and metalinguistic knowledge tests have been identified as measures of implicit and explicit knowledge, respectively, questions still remain as to which of the two constructs grammaticality judgment tests (GJTs) actually measure (R. Ellis, 2005; Isemonger, 2007). This study addresses these questions by examining how manipulations of time pressure and task stimulus affect the performance of learners of Spanish as a second language (L2) on two GJTs.

BACKGROUND

Implicit and Explicit Knowledge of Language

The constructs of implicit and explicit knowledge of the L2 are central to the field of SLA. Issues such as their role in L2 learning and use as well as the relationship between these two types of knowledge are far from settled in the field.[1] In short, although there seems to be a certain degree of agreement in that the ultimate goal of L2 acquisition is the development of implicit knowledge, questions still remain as to what the role of explicit knowledge is. Explicit knowledge may facilitate L2 acquisition by accelerating the establishment of links between form and meaning (R. Ellis, 1994), it may contribute to linguistic problem solving when implicit knowledge is not sufficient (N. Ellis, 2005), or it may be used for the conscious production of output, which, through usage, may lead to implicit learning (N. Ellis, 2011; R. Ellis, 2009). Additionally, although there is ample evidence that indicates that implicit and explicit knowledge are neurologically distinct, it is not clear if they constitute a dichotomy or, rather, if they represent the two ends of a continuum. Likewise, whether or not implicit knowledge representations can become explicit and vice versa (i.e., the interface issue) is also a matter of controversy. However, there seems to be consensus that both types of knowledge interact at the level of performance.

R. Ellis (2005) and Bowles (2011) discuss seven features that distinguish implicit and explicit knowledge: (a) awareness, (b) type of knowledge, (c) systematicity, (d) accessibility, (e) use of knowledge, (f) self-report, and (g) learnability. According to these authors, implicit knowledge is intuitive and procedural, variable but systematic, usually accessed by means of automatic processing and during fluent performance, and not verbalizable; there also seem to be age-related constraints regarding learners' ability to acquire it. Conversely, explicit knowledge is conscious and declarative, often anomalous and inconsistent, and is accessed

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

through controlled processing; it is a tool to achieve control in linguistic problem solving, is potentially verbalizable, and can be learned at any age. On the basis of these seven features, R. Ellis (2005) developed seven criteria on which implicit and explicit knowledge can be operationalized and that constitute the basis for designing tests to measure them.[2] These criteria are:

1. *Degree of awareness.* In tests intended to measure implicit knowledge, learners would resort to their intuitions (i.e., feel), whereas they would access their conscious knowledge of rules in tests of explicit knowledge.
2. *Time available.* Tests that place time constraints on learners would tap into implicit knowledge, whereas tests with no time limits would tap into explicit knowledge.
3. *Focus of attention.* If the focus of the test is on meaning, learners are likely to resort to their implicit knowledge, whereas they would resort to their explicit knowledge when focus is on form.
4. *Systematicity.* Learners' responses are likely to be more consistent in a test that taps implicit knowledge than in an explicit knowledge test.
5. *Certainty.* It is possible that learners would be more certain about the responses to a test for which they resorted to implicit knowledge than those of an explicit knowledge test.
6. *Metalanguage.* Learners' may make use of their knowledge of technical terms in an explicit knowledge test but not in an implicit knowledge test.
7. *Learnability.* Learners who started their L2 learning as a child would likely show higher levels of implicit knowledge than those who started at an older age.

Thus, combining several of these criteria in the design of a particular test would likely make such a test a measure of implicit or explicit knowledge. It is important to note, however, that it is likely impossible to construct tests that are pure measures of one or the other type of knowledge. As R. Ellis (2005) notes, these criteria might predispose learners to draw on one type of knowledge rather than the other, but there is no guarantee that learners would actually do so.

Past Research on GJTs

For decades, GJTs have been used in SLA as a measure of learners' linguistic ability in the L2 (e.g., Bialystok, 1979; Bley-Vroman, Felix, & Ioup, 1988; De Jong, 2005; Han & Ellis, 1998; Loewen & Erlam, 2006; Montrul, 2005; Sorace, 1985; Toth, 2006), and have been popular for a variety of practical and theoretical reasons. The main practical reason for using GJTs is that "some phenomena are not accessible to investigation in production data because they occur either rarely or not at all"

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

(R. Ellis, 1991, p. 163). In other words, it is difficult to test L2 learners' knowledge of certain structures via production because of their low frequency of occurrence in their L2 output, and, in these cases, GJTs can help gather information about learners' knowledge of these structures. Additionally, GJTs are relatively easy to administer to large numbers of learners at once. Regarding theoretical reasons, as Loewen (2009) noted, many researchers (e.g., Bard, Robertson, & Sorace, 1996; Birdsong, 1989; Hedgcock, 1993) agree that "GJTs provide a performance measure of L2 learners' linguistic abilities" (p. 95) through which inferences about linguistic competence can be made.

Grammaticality judgment tests usually consist of a number of grammatical and ungrammatical sentences, and learners are asked to determine whether those sentences are well formed or deviant (R. Ellis, 1991). Additionally, learners may also be asked to do one or several of the following tasks with respect to the ungrammatical sentences in the test: (a) to indicate the error, (b) to correct the error, or (c) to explain the rule violated. According to R. Ellis (2004), GJTs in which learners have to discriminate between grammatical and ungrammatical sentences can be carried out on the basis of either implicit or explicit knowledge, whereas tests that require learners to locate and correct errors likely entail access to explicit knowledge. Finally, tests that require learners to describe rules being violated can only be done on the basis of explicit knowledge. In relation to these tasks, R. Ellis (2004) discussed three principal processing operations in which learners may engage when carrying out a GJT: (a) semantic processing (i.e., "understanding the meaning of a sentence" [p. 256]); (b) noticing (i.e., deciding whether or not there is something ungrammatical in a sentence); and (c) reflecting (i.e., identifying what is incorrect and possibly determining why it is incorrect). Semantic processing and noticing can be carried out using implicit knowledge (R. Ellis, 2009; Ellis & Loewen, 2007), whereas reflecting likely requires access to explicit knowledge as learners try to verify whether or not the sentence is grammatical. Thus, semantic processing and noticing are necessarily involved in determining whether or not a particular sentence is grammatical, whereas reflecting may or may not take place.[3] Consequently, in GJTs in which the learners are only asked to determine the grammaticality of sentences, "the important issue is to determine the nature of the linguistic knowledge, whether implicit, explicit or some combination of both, that participants draw on in their performance on GJTs" (Loewen, 2009, p. 95).

Two aspects of GJTs that have been examined in relation to how they affect learners' judgments in terms of using implicit or explicit knowledge are time pressure and task stimulus. Regarding time pressure, GJTs that do not place any time constraints on learners' to respond allow participants to engage in the three processing operations previously discussed (i.e., semantic processing, noticing, and reflecting). Thus, learners have

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

the opportunity to resort to their explicit knowledge of the L2, although there is no guarantee that they may do so (R. Ellis, 2004). In contrast, GJTs that impose time limits for responding to the test items arguably constrain the learners to accessing their implicit knowledge because the possibility of accessing their explicit knowledge is highly reduced due to the speeded nature of the test (Bialystok, 1979; R. Ellis, 2004). Time pressure, however, is not a guarantee that explicit knowledge cannot be retrieved, particularly if this knowledge has been automated (see DeKeyser, 2003, 2009). Many studies that have used GJTs have allowed unlimited time to respond. However, a few recent studies have employed two versions of a GJT, one with time constraints and one without (Bowles, 2011; R. Ellis, 2005; Han, 2000; Han & Ellis, 1998; Loewen, 2009; Mandell, 1999). The results in these studies have been interpreted as evidence that timed GJTs tap into learners' implicit knowledge, whereas untimed GJTs constitute a measure of explicit knowledge.

With respect to task stimulus, the question remains as to whether or not judging grammatical versus ungrammatical sentences draws on different types of knowledge. The findings of studies that have investigated this aspect are somewhat inconclusive. In a study about the factors that affect the use of explicit and implicit knowledge in a GJT, Bialystok (1979) found that learners made decisions on the well-formedness of the sentences on the basis of their intuitions (i.e., implicit knowledge) regardless of whether the sentences were grammatical or ungrammatical, and regardless of whether the test was timed or untimed. However, she found that learners experienced more difficulty in judging ungrammatical sentences than grammatical ones in the timed test but judged both types of sentences with equal accuracy in the untimed test. Bialystok interpreted this finding as evidence that availability of explicit knowledge and of time facilitate judging ungrammatical sentences. In a later study, Bialystok (1986) also found that ungrammatical sentences are more difficult to judge than grammatical ones, which she attributed to higher demands in control and analysis.

In the analysis of the GJT used in his study, R. Ellis (1991) found that L2 learners judged grammatical sentences more accurately than ungrammatical ones. He noted, however, that "different tests are likely to produce different results depending on the structure under investigation and the proficiency of the learners" (R. Ellis, 1991, p. 176). Similarly, Martens (1988; cited in Hedgcock, 1993) provided further evidence that L2 learners did better at judging grammatical sentences than ungrammatical ones. Bley-Vroman et al. (1988), in contrast, found that the L2 learners in their study judged ungrammatical sentences significantly more accurately than grammatical ones and noted that their L2 subjects "tended to reject more often than they should have" (p. 25). In a study with learners of L2 English at two levels of proficiency, Gass (1983)

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

found that the intermediate group judged grammatical sentences more accurately than ungrammatical ones, whereas the advanced group showed the opposite behavior. In a review article about the differences in judging grammatical and ungrammatical sentences, Hedgcock (1993) acknowledged the plausibility of learners resorting to different types of knowledge when carrying out such judgments by stating that "positing autonomous L2 knowledge systems in which certain tasks or stimuli (i.e., well-formed vs. ill-formed target language utterances) induce learners to tap into separate data modules is, indeed, an attractive way of accounting for variable performance across learners and across tasks" (p. 15). Furthermore, he also emphasized the fact that the design of the test may play a significant role in the subjects' performance.

In one of the most complete studies to date to explore the differential effects of time pressure and task stimulus in GJTs, Loewen (2009) examined the performance of native and L2 English speakers on a timed and an untimed GJT.[4] He found that both the native and the L2 groups judged the grammatical sentences in the timed GJT more accurately than the ungrammatical ones. However, the two groups showed no differences in accuracy in judging the grammatical and ungrammatical sentences in the untimed test. Furthermore, contrary to studies that suggest that learners take longer in judging ungrammatical sentences than grammatical ones (e.g., Juffs, 2001), Loewen found that the L2 learners in his study judged ungrammatical sentences slightly faster in the timed test and significantly faster in the untimed test than they judged grammatical sentences. Although he did not discuss the differences between the L2 learners' scores on grammatical and ungrammatical items on both GJTs in terms of the implicit and explicit knowledge constructs, his interpretation seems to point in this direction. Loewen explained his findings on the basis of the three processes previously discussed (i.e., semantic processing, noticing, and reflecting). In his view, learners engage in semantic processing and noticing to determine whether or not a sentence is grammatical. If it is grammatical, they may make a judgment, but if it is not, they may engage in reflecting to determine what is ungrammatical in the sentence. In untimed GJTs, learners have time for these three processes, and this fact explains the lack of differences between their judgments of grammatical and ungrammatical sentences according to Loewen. However, in timed GJTs, learners only have time for semantic processing and noticing, not for reflecting. Therefore, it is reasonable to expect to find significant differences between learners' judgments of grammatical and ungrammatical sentences, as was the case in Loewen's study. If we agree with R. Ellis (2009) and Ellis and Loewen (2007) that learners are able to determine whether or not sentences are grammatical on the basis of their implicit knowledge, it thus follows that semantic processing and noticing can be carried out on the basis of this type of knowledge. Additionally, learners need to

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

resort to their explicit knowledge to engage in reflecting. Consequently, in timed GJTs, learners have enough time to access their implicit knowledge to judge grammatical sentences but not to access their explicit knowledge to judge ungrammatical ones. When GJTs do not contain any time constraints, learners have time to access not only their implicit representations to judge grammatical sentences but also their explicit knowledge to judge ungrammatical sentences.

Several studies have examined the type of knowledge representations on which L2 learners draw when completing GJTs. Davies and Kaplan (1998), R. Ellis (1991), and Goss, Ying-Hua, and Lantolf (1994) analyzed the learners' verbalizations during GJTs in an attempt to identify the strategies they used when judging sentences. All three studies found that learners resorted to both their implicit and explicit knowledge of the L2, among other strategies, when judging the sentences. However, they did not provide details as to what aspects of the GJT led the learners to draw on one or the other type of knowledge.

More recent studies have employed factor analysis—either exploratory (R. Ellis, 2005; Han & Ellis, 1998), confirmatory (Bowles, 2011), or both (Ellis & Loewen, 2007[5])—to find out whether or not various tests, including timed and untimed GJTs,[6] are indeed valid measures of implicit and explicit knowledge. In Han and Ellis (1998) and Bowles (2011), the timed GJT was identified as a measure of implicit knowledge and the untimed GJT as a measure of explicit knowledge. These two studies, however, did not examine differences in the learners' performance on grammatical and ungrammatical items in the GJTs. In R. Ellis (2005), the untimed GJT loaded quite strongly on both the implicit and the explicit factor. For this reason, R. Ellis decided to run a second exploratory factor analysis using only the scores for the ungrammatical sentences in the untimed GJT, on the basis of the argument previously discussed that learners likely resort to their implicit knowledge to judge grammatical sentences, whereas they draw on their explicit knowledge to judge ungrammatical ones. In this second factor analysis, the scores on the ungrammatical section of the untimed GJT loaded only on the explicit knowledge factor. In their follow-up study aimed at confirming the findings in R. Ellis (2005), Ellis and Loewen (2007) also used only the scores for the ungrammatical section of the untimed GJT. The factor analyses in both studies were interpreted as evidence that the scores for the ungrammatical items on the untimed test constituted a measure of explicit knowledge. However, these two studies did not examine differences between the two types of task stimuli in the timed test or the scores for the grammatical items on the untimed test in their factor analyses.

R. Ellis (2005, 2009) and Loewen (personal communication, October 30, 2009) expressed their concerns regarding timed GJTs as a valid measure of implicit knowledge. Given the doubts as to what GJTs actually measure,

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

Isemonger (2007) noted that "GJTs are central to a future debate on the psychometrics of implicit and explicit knowledge, especially because so much past research has been based on them" (p. 110). To date, however, no study has used factor analysis to empirically examine if grammatical and ungrammatical sections of GJTs load on the same or different factors. R. Ellis (personal communication, March 28, 2011) suggested carrying out such a study by separating learners' scores on a timed and an untimed GJT between grammatical and ungrammatical sentences and conducting exploratory and confirmatory factor analyses to test whether the former constitute a measure of implicit knowledge and the latter a measure of explicit knowledge; the present study aims to do just this.

THE STUDY

The study reported here is part of a larger study about the nature of the knowledge representations developed by learners of L2 Spanish. The purpose of the study was to examine whether or not L2 learners draw on different types of knowledge in judging grammatical and ungrammatical sentences in timed and untimed GJTs. The following research questions guided this study:

1. a) Do L2 learners' responses to grammatical sentences in timed and untimed GJTs differ from those to ungrammatical sentences in such tests?
   b) If so, can these differences be interpreted in terms of the implicit versus explicit knowledge distinction?
2. How do time pressure and task stimulus affect the L2 learners' performance on GJTs?

Participants

The study took place at a Canadian university and involved 49 participants (female: $n = 40$; male: $n = 9$). The participants were enrolled in one of two university-level courses of L2 Spanish and had returned a signed consent form. Twenty-nine of the participants were near the end of their third term of Spanish language instruction at the university; this level of instruction is intended to help learners develop their proficiency to a level similar to the A2 level of the Common European Framework of Reference (CEFR).[7] The other 20 participants were close to completing their fifth term of Spanish language training; the goal is for learners to have reached a level similar to the B1 level of the CEFR at this point in

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

their L2 development. All the participants were nonnative speakers of Spanish. Twenty-two participants had taken Spanish language courses prior to university, most of them in high school. Forty-one participants started Spanish language training at the beginners' level at university, including 14 of the 22 who had taken Spanish previously but who did not meet the proficiency requirements of the intermediate level. The remaining eight participants had been placed in intermediate-level courses. The majority of participants described the type of instruction they had received at university or at other institutions as either mainly formal (i.e., with a focus on explicit grammar instruction) or a mixture of formal and informal (i.e., with a focus on explicit grammar instruction combined with communicative practice). On average, the participants reported having studied Spanish for 2.1 years. Twenty-four of the participants reported having lived in or visited a Spanish-speaking country, 17 of them for less than a month, and seven between 2–5 months.

Instruments and Procedures

The participants completed a language background questionnaire, a timed GJT, an untimed GJT, and a metalinguistic knowledge test (MKT). The language background questionnaire was employed to obtain personal information about the participants and some details about their exposure to Spanish (see previous section). The two GJTs and the MKT were designed following the criteria in R. Ellis (2005). Thus, the timed GJT was intended to encourage learners to use their intuitions; it imposed time constraints, and, thus, even though the focus was on form, learners should have had little opportunity to access their metalinguistic knowledge. Consequently, it was predicted that this test would tap into learners' implicit knowledge of the L2. Conversely, the untimed GJT and the MKT encouraged the use of knowledge of rules; they did not impose time constraints and, thus, should have allowed learners the opportunity to access their knowledge of technical metalanguage. On the basis of these design features, it was believed that these two tests would measure learners' explicit knowledge.

Both GJTs contained 64 sentences, half of which were grammatical and half of which were ungrammatical, and the two GJTs were administered in two separate sessions. The 64 sentences covered 16 grammatical structures (see Appendix) that had been explicitly taught in the university courses that the participants had taken or were taking at the time of the study.[8] These structures were selected because they have been identified as being typically problematic for L2 learners of Spanish even after years of instruction (Bowles, 2011; Montrul, 2011). For example, the different uses of (a) *ser* and *estar*, (b) the preterit and imperfect past

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

tenses, and (c) the subjunctive are challenging for many learners even at advanced levels. Another reason for choosing these structures is that they represent grammar aspects introduced at different instructional points throughout the curriculum. For instance, determiner-noun agreement and subject-verb agreement are introduced early at the beginner level, whereas the uses of the subjunctive are introduced at the intermediate level. For each of the grammatical structures, there were two grammatical sentences and two ungrammatical ones. The sentences were designed to contain vocabulary with which the learners were familiar so as to maximize understanding, although there is no guarantee that all learners actually understood each and every sentence. The sentences were the same in both the timed and untimed GJT.

For the timed GJT, the sentences were presented on an automated PowerPoint slide show, and the participants were asked to write their responses on an answer sheet.[9] Sentences remained on the screen 6–9 s depending on sentence length. These times were estimated on the basis of previous studies in which sentences were shown for 3 or 3.5 s (Bialystok, 1979; Han, 2000) or up to 6.24 s (Loewen, 2009). Therefore, depending on sentence length, it was decided to allow 3–6 s for sentence processing, and an additional 3 s for writing the response on the answer sheet. An auditory cue signaled the transition between sentences. At the beginning of the test, the participants were instructed to indicate whether the sentences that they were about to be shown were grammatical or ungrammatical, writing "G" or "U" on the answer sheet.[10] The participants were also reminded that there was a time limit for each sentence and that they would need to respond as quickly as possible. They were also asked to respond to eight practice sentences so that they would get used to the speeded nature of the test. Between each practice sentence a reminder of the time limit for judging the sentences was shown. Participants were given a 15-s break every 16 sentences.

The untimed GJT, as previously noted, contained the same 64 sentences as the timed version. The participants received an answer sheet and a separate sheet containing the sentences, and were instructed to indicate whether the sentences were grammatical or ungrammatical on the answer sheet. The participants were reminded that there was no time limit to complete the test.

The MKT consisted of 16 sentences covering the same grammatical structures as the GJTs. Each sentence contained an underlined error, and the participants were asked to provide a written explanation in English of the rule violated in each of the sentences. This test was administered during a third session, and there was no time limit to complete it. It must be emphasized that the main focus of this study is the results obtained on the two GJTs. However, the MKT has been included in the present study because it has been identified as a measure of explicit knowledge in previous studies (Bowles, 2011; R. Ellis, 2005;

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

Ellis & Loewen, 2007). As such, it was hypothesized that any other measures used in this study that loaded on the same factor as the MKT in the factor analyses could also be considered measures of explicit knowledge.

Scoring Procedures

Both GJTs were scored in terms of correct versus incorrect answers. Although items left blank could be interpreted as (a) the participants not being sure about the grammaticality of the sentence or (b) the participants not having had enough time to judge the sentence, a decision was made to count such an answer as incorrect for purposes of comparability with R. Ellis (2005) and Bowles (2011). Furthermore, blank items accounted for a very small percentage of all the responses to the timed and untimed GJTs: 7.33% and 1.83%, respectively. Correct answers were awarded 1 point, whereas incorrect answers were awarded 0 points. For both tests, the scores were calculated for the test as a whole (i.e., 64 points maximum) and also separately for grammatical and ungrammatical sentences (i.e., 32 points maximum in each section).

The MKT was also scored in terms of correct and incorrect answers for purposes of comparability. Answers were considered correct if a rule was formulated adequately regardless of the use of technical metalanguage. Imprecise or incomplete rule formulations were considered incorrect. Eighteen tests (close to 37%) were randomly selected and scored by a second rater. Interrater agreement was 89.93%, and all disagreements were resolved through discussion.

Analysis

The reliability of the three tests was calculated using Cronbach's alpha. Descriptive statistics were computed for the total scores on the timed and untimed GJTs, for the scores on the grammatical and ungrammatical sentences in both GJTs, and for the scores on the MKT. Additionally, Pearson product moment coefficients were calculated to examine the correlations between the five measures. A principal component factor analysis (performed in SPSS 19) with oblique rotation and two confirmatory factor analyses (performed in AMOS 19; Arbuckle, 2010) were carried out to examine Research Question 1.[11] To address Research Question 2, a repeated-measures ANOVA and a paired-samples *t* test were carried out to examine the main effects of time pressure and task stimulus on the two GJTs as well as the interactions and differences

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

between these two dimensions. To gain further insight into the learners' performance on the GJTs, an item analysis that examined learners' responses for the different types of grammatical structures included in the test was carried out. All statistical tests are described in more detail in the **RESULTS** section.

RESULTS

Table 1 shows the measure of reliability for the timed and untimed GJT, and for the MKT. Cronbach's alpha coefficient was .78 for the timed test, .83 for the untimed one, and .82 for the MKT, indicating that the three tests were reliable (Field, 2009).

Table 2 shows the mean and standard deviation for the total scores on the two GJTs, for the scores on the grammatical and ungrammatical sentences in those tests, and for the scores on the MKT. With respect to the total scores in the two GJTs, the participants obtained higher scores in the untimed test than in the timed one. The standard deviations show a similar variance in both tests. Regarding the task stimulus distinction, the participants scored considerably higher on the grammatical sentences than on the ungrammatical ones in both the timed and untimed test. Indeed, the learners did quite poorly in judging the ungrammatical sentences, particularly in the timed test. Additionally, the standard deviations show greater variance in the ungrammatical sentences than in the grammatical ones. With respect to the MKT, the results show that the participants also did quite poorly and that the scores varied considerably.

Table 3 presents the correlations matrix between the grammatical and ungrammatical sentences in both the timed and untimed GJT and the MKT. As the table shows, all four sections of the GJTs are strongly correlated. However, the results show larger effect sizes for the correlation between grammatical sentences in the timed and untimed tests (r = .76) and for that between the ungrammatical sentences in the timed and untimed tests (r = .70) than for the correlation between grammatical and ungrammatical sentences in the same test or across tests. With respect to the MKT, the grammatical and ungrammatical sections in both the timed and untimed GJTs correlated significantly with the MKT

Table 1. Reliability coefficients for the three tests

Test	Items	Participants	Reliability Coefficient
Timed GJT	64	49	α = .78
Untimed GJT	64	49	α = .83
MKT	16	49	α = .82

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

Table 2. Descriptive statistics for the tests (N = 49)

Test and Items	*M* correct (%)	*SD*
Timed GJT total	57.78	11.45
Timed GJT grammatical	71.62	10.52
Timed GJT ungrammatical	43.94	16.66
Untimed GJT total	66.04	12.74
Untimed GJT grammatical	77.17	11.70
Untimed GJT ungrammatical	55.52	18.67
MKT	53.76	21.88

scores, although the correlations with the ungrammatical sections show larger effect sizes than the ones with the grammatical sections.

A principal components factor analysis was conducted to examine whether or not grammatical sentences in both the timed and untimed GJT could be considered measures of implicit knowledge and, conversely, whether or not ungrammatical sentences in both tests constitute measures of explicit knowledge. The initial analysis, presented in Table 4, showed that two components had eigenvalues greater than 1 (Kaiser, 1960). These two components combined accounted for 79.01% of the total variance.

Table 5 shows the results of the principal components factor analysis. Following the recommendations of Field (2009) and Isemonger (2007), an oblique rotation (direct oblimin) was chosen on the basis of the assumption that the factors would be correlated. As the table shows, the ungrammatical sections of both the timed and untimed GJT and the MKT load heavily (above .7) on Factor 1, identified as explicit knowledge, whereas the grammatical sections of both tests load heavily (above .9) on Factor 2, identified as implicit knowledge.

In his critique of R. Ellis's (2005) study, Isemonger (2007) pointed out that, given that the purpose of the study was to verify an already

Table 3. Correlation matrix for the grammatical and ungrammatical sections in both GJTs, and for the MKT (N = 49)

Test	TGJT Gr	TGJT Un	UGJT Gr	UGJT Un	MKT
TGJT Gr	—	.389**	.755**	.393**	.318*
TGJT Un		—	.355*	.704**	.456**
UGJT Gr			—	.359*	.395**
UGJT Un				—	.622**
MKT					—

Note. * $p < .05$ (2-tailed); ** $p < .001$ (2-tailed); TGJT = timed GJT; UGJT = untimed GJT; Gr = grammatical; Un = ungrammatical.

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

Table 4. Principal component factor analysis

Component	Eigenvalue	Variance (%)	Cumulative (%)
1	2.903	58.054	58.054
2	1.048	20.955	79.009

established theoretical model, confirmatory factor analysis (CFA) should have been used instead of exploratory factor analysis. Isemonger argues that exploratory factor analysis does not allow testing rival hypotheses or falsifying models. Following this recommendation, Ellis and Loewen (2007) and Bowles (2011) conducted confirmatory factor analyses to examine whether the tests used could be considered measures of implicit and explicit knowledge. For the present study, two confirmatory factor analyses were conducted using AMOS 19 (Arbuckle, 2010). The first analysis tested the construct solution predicted in the exploratory factor analysis previously discussed—namely, that the grammatical sections of the timed and untimed GJTs would load on an implicit factor and that the ungrammatical sections of both tests and the MKT would load on an explicit factor (Grammatical/Ungrammatical Model). The second analysis tested an alternative model, as recommended by Isemonger (2007). In this case, the alternative model was that obtained in the first principal components analysis in R. Ellis (2005) and confirmed by Bowles' (2011) CFA—namely, that the timed GJT (grammatical and ungrammatical sections) would load on an implicit factor, and that the untimed GJT (grammatical and ungrammatical sections) and the MKT would load on an explicit factor (Timed/Untimed Model). The results of these analyses are presented in Figures 1 and 2, respectively.

A summary of the model of fit for both solutions is presented in Table 6. The chi-square statistic was used to test the absolute goodness of fit of the model. A nonsignificant value indicates an acceptable fit of the model, whereas a significant value indicates that the model is statistically unlikely to occur (Byrne, 2001). As shown in Table 6, the Grammatical/

Table 5. Loadings for principal components factor analysis

Test	Component 1	Component 2
UGJT Un	.933	-.034
TGJT Un	.842	.008
MKT	.797	.025
TGJT Gr	.000	.938
UGJT Gr	.001	.935

Note. Rotation method = Direct oblimin; UGJT = untimed GJT; TGJT = timed GJT; Un = ungrammatical; Gr = grammatical.

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

Ungrammatical Model (i.e., the model in which the grammatical sections load on the implicit factor and the ungrammatical sections load on the explicit factor) is an acceptable fit, whereas the Timed/Untimed Model (i.e., the model in which the timed GJT loads on the implicit factor and the untimed GJT loads on the explicit factor) is not. Additionally, four other statistics of fit were used to assess the overall fit of the model. Values of 0.95 or higher in the normed fit index (NFI) and in the Tucker-Lewis index (TLI) indicate a superior fit for the model. A value of .05 or lower in the root mean square error of approximation (RMSEA) indicates a good fit for the model, whereas models with an RMSEA of .10 or higher have a poor fit. Finally, significance of close fit (PCLOSE) values higher than .05 indicate a close fit. The figures reported in Table 6 show that the criteria for these four statistics of fit are met for the Grammatical/Ungrammatical Model, but not for the Timed/Untimed Model.

To address Research Question 2, a repeated-measures ANOVA with time pressure and task stimulus as independent variables was computed to examine the effect of these two variables on the learners' performance on the GJTs. The results in Table 7 show that there was a significant effect for time pressure, $F(1, 48) = 57.50$, $p < .001$, $r = .74$, indicating that

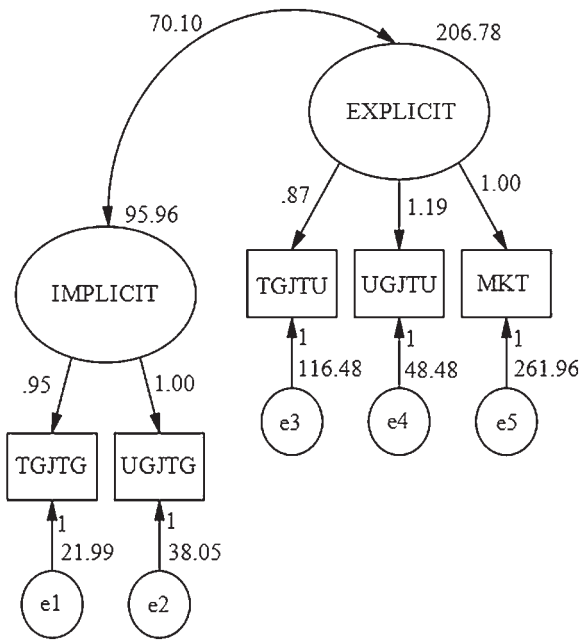


Figure 1. Grammatical/ungrammatical model. TGJTG = timed GJT, grammatical section; UGJTG = untimed GJT, grammatical section; TGJTU = timed GJT, ungrammatical section; UGJTU = untimed GJT, ungrammatical section; MKT = metalinguistic knowledge test.

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

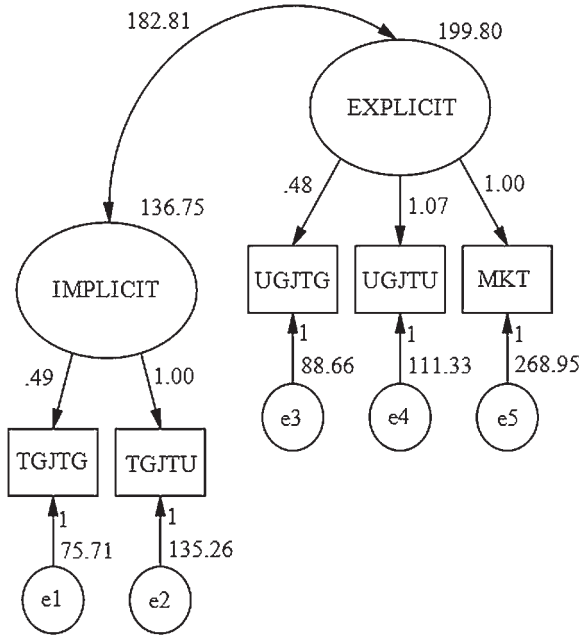


Figure 2. Timed/untimed model. TGJTG = timed GJT, grammatical section; TGJTU = timed GJT, ungrammatical section; UGJTG = untimed GJT, grammatical section; UGJTU = untimed GJT, ungrammatical section; MKT = metalinguistic knowledge test.

the learners' scores were significantly more accurate on the untimed GJT than on the timed one. Furthermore, the results show a significant effect for task stimulus, $F(1, 48) = 130.82$, $p < .001$, $r = .86$, indicating that the learners' scores were significantly more accurate on the grammatical sections of the GJTs than on the ungrammatical ones. Finally, the table also shows a significant interaction at the $p < .05$ level with a medium effect size ($r = .36$) between time pressure and task stimulus. This interaction is illustrated in Figure 3. The differences between the grammatical sentences (solid line) and the ungrammatical ones (dashed line) are quite similar on both the timed and untimed GJTs. Indeed, an examination of the

Table 6. Summary of the model of fit for the solutions in figures 1 and 2

Model	χ^2	NFI	TLI	RMSEA	PCLOSE	*df*
Grammatical/ Ungrammatical	3.500	0.968	1.012	.000	.539	4
Timed/Untimed	33.466**	0.696	0.265	.392	.000	4

Note. ** $p < .00$

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

Table 7. Repeated-measures ANOVA

Variable	*df*	*F*	*p*	*r*
Time Pressure	1	57.498	.000	.74
Grammaticality	1	130.817	.000	.86
Time Pressure × Grammaticality	1	7.168	.010	.36

means reported in Table 2 shows that the difference between grammatical and ungrammatical scores was 27.68% on the timed test and 21.65% on the untimed one. As the figure shows, the learners had more difficulties with both the grammatical and the ungrammatical sentences in the timed GJT than in the untimed one. The data presented in Table 2 also show that the scores for the grammatical sentences in the untimed GJT were 5.55% higher than those for the grammatical sentences in the timed test. Likewise, the scores for the ungrammatical sentences in the untimed test were 11.58% higher than the scores in the same section of the timed test.

A paired-samples *t* test was computed to examine in more detail the differences between the scores obtained by the learners in the grammatical and ungrammatical sentences in each of the two tests. Table 8 shows the results of this test. As the table shows, all the differences between the means were significant at the $p < .001$ level and with large effect sizes. It is worth noting that both the mean differences and the effect sizes were larger for the comparison between grammatical and ungrammatical sections than for the comparison of timed versus untimed tests.

An item analysis that examined the learners' responses with respect to each of the grammatical structures tested on the GJTs was carried out to gain further insights into the differences in learners' performance in terms of time pressure and task stimulus. Tables 9 and 10 show the differences between the timed and untimed GJTs and between the grammatical and ungrammatical sentences, respectively, for each of the grammatical structures. As can be seen in Table 9, the learners performed better on the untimed test than on the timed one on all of the structures except two—*ser* or *estar* with personality and imperfect with habitual actions—for which there were no differences. Table 10 shows that learners' responses were more accurate on the grammatical sentences than on the ungrammatical ones for each of the structures except one—subject-verb agreement in future tense—which showed the opposite pattern.

DISCUSSION

The main purpose of this study was to find out whether or not L2 learners' judgments of grammatical sentences in both a timed and an

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

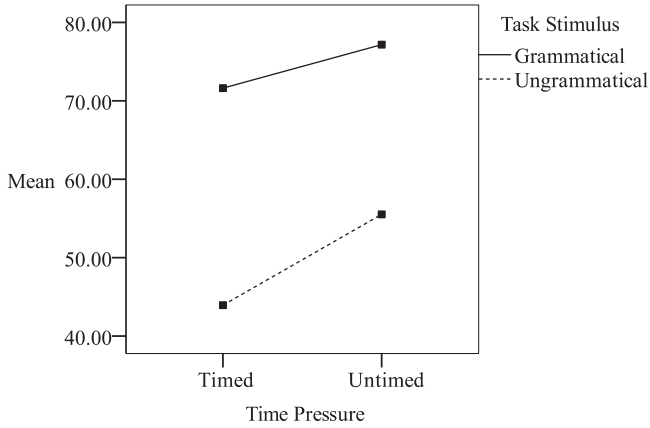


Figure 3. Interaction of time pressure and task stimulus in the GJTs.

untimed GJT can be interpreted as a measure of L2 implicit knowledge, and whether or not judgments of ungrammatical sentences in both tests may be indicative of L2 explicit knowledge (Research Question 1). The study also sought to examine the effects of time pressure and task stimulus on L2 learners' performance on those tests (Research Question 2). The descriptive statistics and the results of the exploratory and the confirmatory factor analyses provide answers to the first research question. The mean accuracy scores on the grammatical items were quite different from those on the ungrammatical ones in both GJTs. Furthermore, both factor analyses showed that the scores on the grammatical sections in both GJTs and those on the ungrammatical ones load on two separate factors. These results indicate that learners in this study processed grammatical and ungrammatical sentences in the GJTs differently. A plausible interpretation of these results is that the grammatical sections of the timed and untimed GJTs likely measure the construct of implicit knowledge, whereas the ungrammatical sections of both GJTs possibly measure the construct of explicit knowledge. A key aspect for this interpretation is that, in both exploratory and confirmatory factor analyses, the scores for the ungrammatical sentences in both the timed and untimed GJTs loaded on the same factor as the MKT, which, as noted, has been identified in previous studies as a measure of explicit knowledge. An alternative confirmatory factor analysis was carried out to examine whether or not the scores could also fit a model that is based on the distinction between timed and untimed tests as in R. Ellis (2005) and Bowles (2011). The results of this second CFA, however, indicated that such a model was not acceptable for the data collected in this study. Additional evidence for the acceptable model comes from the results of the correlations between the different sections in the GJTs and the MKT. Although the scores on the MKT correlated significantly with

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

Table 8. Paired-Samples t test

	Paired Differences				
Test	*M* (%)	*SEM*	*t*	*df*	*r*
TGJT Gr vs. TGJT Un	27.68	2.27	12.21**	48	.87
UGJT Gr vs. UGJT Un	21.65	2.59	8.36**	48	.77
TGJT Gr vs. UGJT Gr	-5.55	1.12	-4.95**	48	.58
TGJT Un vs. UGJT Un	-11.58	1.96	-5.91**	48	.65

Note. ** $p < .001$; TGJT = timed GJT; UGJT = untimed GJT; Gr = grammatical; Un = ungrammatical

those on the grammatical and ungrammatical sentences in the timed and untimed GJT, as Table 3 shows, the correlations with the ungrammatical sections of both GJTs showed a larger effect size than those with the grammatical sections. Furthermore, R. Ellis (2005) argued that measures of implicit knowledge would obtain more systematic (i.e., less variable) responses than measures of explicit knowledge. In line with this argument, the greater variance in the ungrammatical sentences than in the grammatical ones shown in Table 2 can be interpreted as evidence of learners' resorting to explicit knowledge for the ungrammatical sentences and to implicit knowledge for the grammatical sentences.

Table 9. Mean differences between timed and Untimed GJT

	M (%)		
Structure[a]	TGJT	UGJT	Diff (%)
1	57.14	64.80	-7.66
2	79.08	87.24	-8.16
3	75.00	78.57	-3.57
4	59.18	66.33	-7.15
5	59.18	74.49	-15.31
6	51.53	69.39	-17.86
7	48.98	63.78	-14.80
8	66.33	73.98	-7.65
9	61.73	76.53	-14.80
10	49.49	49.49	0.00
11	56.12	61.73	-5.61
12	58.16	69.90	-11.74
13	60.71	60.71	0.00
14	61.22	62.76	-1.54
15	50.00	56.63	-6.63
16	49.49	51.02	-1.53

Note. [a] See appendix for list of structures; TGJT = timed GJT; UGJT = untimed GJT; Diff = difference

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

Table 10. Mean differences between Grammatical and Ungrammatical Sentences

	M (%)		
Structure[a]	Gram	Ungram	Diff (%)
1	88.78	33.16	55.62
2	90.82	75.51	15.31
3	84.18	69.39	14.80
4	59.18	66.33	-7.15
5	84.18	49.49	34.69
6	72.96	47.96	25.00
7	77.04	35.71	41.33
8	84.18	56.12	28.06
9	79.59	58.67	20.92
10	74.49	24.49	50.00
11	69.39	48.47	20.92
12	68.88	59.18	9.70
13	70.41	51.02	19.39
14	76.53	47.45	29.08
15	73.98	32.65	41.33
16	54.59	45.92	8.67

Note. [a] See appendix for list of structures; Gram = grammatical; Ungram = ungrammatical; Diff = difference.

To examine how time pressure and task stimulus affect the L2 learners' performance on the GJTs (Research Question 2), a repeated-measures ANOVA and a paired-samples *t* test were computed, and an item analysis was performed. The results of the repeated-measures ANOVA show that both time pressure and task stimulus have a significant effect on the GJT scores, and that there is a significant interaction between these two variables. These results are similar to those in Loewen (2009). However, Loewen found that the differences between grammatical and ungrammatical sentences were considerable only for the timed test but not for the untimed one, whereas in the present study the differences are considerable in both tests (see Figure 3). The paired-samples *t* test provides additional information about the nature of this interaction. As the results show, the learners in this study judged grammatical sentences significantly more accurately than ungrammatical ones in both the timed and the untimed GJT. Furthermore, they judged both types of sentences significantly more accurately in the untimed test than in the timed one. Thus, these results indicate that both time pressure and task stimulus play an important role in the learners' performance on GJTs. Interestingly, the mean differences between grammatical and ungrammatical sentences are considerably larger (27.68% in the timed GJT and 21.65% in the untimed one) than the mean differences between

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

the timed and the untimed test (-5.55% for grammatical sentences and -11.58% for ungrammatical ones), which indicates that task stimulus seems to have a larger effect than time pressure, as shown by the effect sizes reported in Table 8.

The item analysis of the 16 grammatical structures provides additional information about the learners' performance on the GJTs with respect to time pressure and task stimulus. Regarding time pressure, the results show that learners' responses were better on the untimed test than on the timed one for all grammatical structures except two, which showed no difference. Furthermore, the differences ranged from 1.54% to 17.86%, although it must be noted that these differences were greater than 10% in only five of the 16 structures tested. With respect to task stimulus, the learners did better on the grammatical sentences than on the ungrammatical sentences for all the structures except one—subject-verb agreement in future tense—which obtained a higher average on the ungrammatical sentences. Additionally, the differences between the structures ranged from 7.15%, precisely in subject-verb agreement in future, to 55.62%. For task stimulus, the differences were greater than 15% in 12 of the 16 grammatical structures. These results provide further evidence that different task stimuli had a stronger effect on learners' performance on the GJTs than time pressure.

LIMITATIONS AND CONCLUSIONS

The study reported here is not exempt from limitations. For example, the strength of confirmatory factor analysis benefits from having as many diverse measures as possible for each construct in the model (Isemonger, 2007). Generally, in a two-factor model at least two measures per factor are required, although three measures per factor are usually recommended: "Two *might* be fine, three is better, four is best, and anything more is gravy" (Kenny, 1979, p. 143; see also Klein, 2005); the present study includes two measures of implicit knowledge and three measures of explicit knowledge. This limitation, however, is general to other studies that use CFA to examine the constructs of implicit and explicit knowledge.[12] Therefore, an examination of whether different task stimuli load on different factors would benefit from separating the grammatical and ungrammatical sections of a timed and an untimed GJT and including them within a model that also includes tests such as those used by R. Ellis (2005) and Ellis and Loewen (2007), and by Bowles (2011), which was not possible in the present study due to the limited resources available. Indeed, splitting the scores on both GJTs into grammatical and ungrammatical would yield four measures for the implicit knowledge construct (i.e., oral production task, elicited imitation task, grammatical section in timed GJT, and grammatical section in

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

untimed GJT) and three measures for the explicit knowledge construct (i.e., ungrammatical section in timed GJT, ungrammatical section in untimed GJT, and MKT) and, hence, it would overcome this limitation.

Another limitation regarding the use of factor analysis has to do with the relatively small sample size of this study. Most of the literature (e.g., Cattell, 1978; Field, 2009; Gorsuch, 1983; Kline, 1994; Mulaik, 2009; Warner, 2008) provides recommendations that are based on "the minimum necessary sample size, N, or the minimum ratio of N to the number of variables being analyzed, p" (MacCallum, Widaman, Zhang, & Hong, 1999, p. 84). Thus, recommendations range from the minimum N being between 100 and 500 subjects, or a ratio of N equal to or larger than a range from 3 to 10 times p. If the 100–500 subjects criterion is taken, the sample size used in this study would obviously be too small, whereas if the 3–10 times p criterion is used, the sample size of 49 would be acceptable (or close to) because it represents a ratio of $N > 9p$. It must be pointed out, however, that some studies have demonstrated that relatively small samples may be adequate under certain conditions and have consequently noted that these recommendations should not be taken as rules for every study that uses factor analysis. For example, Barrett and Kline (1981), Arrindell and van der Ende (1985), and MacCallum et al. (1999) found that good recovery of full-sample solutions could be obtained with small subsamples, even when N is well below 100. Furthermore, Pennell (1968), Velicer, Peacock, and Jackson (1982), Velicer and Fava (1998), and MacCallum et al. (1999) found that the necessary sample size depends mainly on the level of communality of the variables and, to a lesser extent, on the level of overdetermination of the factors. In this sense, MacCallum et al. (1999) argue that the impact of the sample size is greatly reduced when the mean level of communality is .70 or higher and all communalities are consistently high (greater than .60). In the present study, the mean level of communality is .79, and the lowest communality is .64. In any case, given the discrepancies in the literature regarding adequate sample size for factor analysis, the results of the present study need to be taken with certain caution.

Despite these limitations, the results of this study indicate that there were processing differences between grammatical and ungrammatical sentences in the two GJTs used, as shown by the significant differences in performance on the two types of sentences. These differences may be interpreted as these two types of sentences constituting measures of implicit and explicit knowledge, respectively. Additionally, the results show that the two dimensions of time pressure and task stimulus significantly affect the L2 learners' performance on those tests. These findings have important implications for the use of GJTs in SLA research and support R. Ellis's (2009) suggestion that "SLA researchers need to take great care to distinguish these two properties in both the design of GJTs

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

and the analysis of scores obtained from tests, as they will influence what is being measured" (p. 60). As Loewen (2009) already pointed out, manipulating the design features of GJTs in terms of presence or absence of time constraints and in terms of judging grammatical or ungrammatical sentences "may predispose L2 learners to draw on different types of linguistic knowledge in making their judgments" (p. 111). Obviously, there is need for future research regarding the effects of time pressure and task stimulus on the learners' performance on GJTs. Nevertheless, the study reported here clearly highlights the need for analyzing the scores on grammatical and ungrammatical sentences in timed and untimed GJTs separately in future studies about implicit and explicit L2 knowledge. More important, the results of the study also emphasize the need to verify whether or not the instruments used in a study actually measure what they are intended to measure (Douglas, 2001), as the construct validity of any testing instrument hinges not only on aspects of the instrument—such as the linguistic structures covered in the test—but also on other aspects such the proficiency of the learners, the length of L2 exposure, and the type of instruction (R. Ellis, 1991; Hedgcock, 1993). Overall, the results of this study make a potentially significant contribution regarding the operationalization and measurement of the learners' implicit and explicit knowledge of the L2, an issue of central importance to the field of SLA (Bowles, 2011; R. Ellis, 2005, 2009; Hulstjin, 2005).

Received 5 August 2011
Accepted 3 November 2011
Final Version Received 22 May 2012

NOTES

1. A detailed discussion of these issues is beyond the scope of this paper. For such a discussion, the reader is referred to DeKeyser (2003, 2009), N. Ellis (2005, 2011), and R. Ellis (2004, 2005, 2009).

2. R. Ellis (2005) designed the tests in his study on the basis of four of these seven criteria: degree of awareness, time availability, focus of attention, and metalanguage.

3. As noted by an anonymous *SSLA* reviewer, this is not to say that these three operations take place in a sequential order or that they are tidily separated in actual processing. For example, semantic processing likely entails some degree of morphosyntactic decoding. Likewise, it is possible to engage in reflecting about the grammaticality of a particular sentence without resorting to explicit knowledge.

4. Loewen's (2009) study is a more detailed analysis of the GJTs reported on in R. Ellis (2005).

5. The participants and tests in Ellis and Loewen (2007) are the same as in R. Ellis (2005). In the 2007 study, following Isemonger's (2007) recommendations, Ellis and Loewen carried out an exploratory factor analysis with a different rotation method and a confirmatory factor analysis.

6. The other tests examined were an oral production task, an elicited imitation task, and a metalinguistic knowledge test.

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

7. In the CEFR, A1 is the lowest level, whereas C2 is the highest. Obviously, there is no guarantee that all learners would actually reach the A2 level or the B1 level after three and five terms of exposure.

8. A complete copy of the GJTs may be provided upon request.

9. Unfortunately, due to the limited resources available for this study, it was not possible to use a computerized test as in R. Ellis (2005) or Bowles (2011).

10. An anonymous *SSLA* reviewer noted that potentially more nuanced data would have been obtained if a Likert-type scale had been used to rate grammaticality instead of the binary choice between grammatical and ungrammatical. This point is well taken. However, this binary choice was adopted for purposes of comparability with previous studies (Bowles, 2011; R. Ellis, 2005). See Tremblay (2005) for a detailed discussion regarding advantages and shortcomings of both types of responses.

11. These statistical procedures were used for purposes of comparability with previous studies (Bowles, 2011; Ellis & Loewen, 2007). It must be noted, however, that there are certain limitations to the use of factor analysis in this study that necessitate exercising caution when considering the results, as will be discussed.

12. Ellis and Loewen (2007) and Bowles (2011) also had two and three measures per construct.

REFERENCES

Arbuckle, J. (2010). AMOS (Version 19.0) [Computer Program]. Chicago: SPSS.

Arrindell, W. A., & van der Ende, J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement*, *9*, 165–178.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*, 32–68.

Barrett, P. T., & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality Study in Group Behavior*, *1*, 23–33.

Bialystok, E. (1979). Explicit and implicit judgments of L2 grammaticality. *Language Learning*, *29*, 81–103.

Bialystok, E. (1986). Factors in the growth of linguistic awareness. *Child Development*, *57*, 498–510.

Birdsong, D. (1989). *Metalinguistic performance and interlinguistic competence*. New York: Springer Verlag.

Bley-Vroman, R., Felix, S., & Ioup, G. (1988). The accessibility of Universal Grammar in adult language learning. *Second Language Research*, *4*, 1–32.

Bowles, M. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in Second Language Acquisition*, *33*, 247–271.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Cattell, R. B. (1978). *The scientific use of factor analysis*. New York: Plenum.

Davies, W. D., & Kaplan, T. (1998). Native speaker vs. L2 learner grammaticality judgments. *Applied Linguistics*, *19*, 183–203.

De Jong, N. (2005). Can second language grammar be learned through listening? An experimental study. *Studies in Second Language Acquisition*, *27*, 205–234.

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). Oxford: 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. 119–138). Oxford: Blackwell.

Douglas, D. (2001). Performance consistency in second language acquisition and language testing: a conceptual gap. *Second Language Research*, *17*, 442–456.

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

Ellis, N. (2011). Implicit and explicit SLA and their interface. In C. Sanz, & R. P. Leow (Eds.), *Implicit and explicit language learning: Conditions, processes, and knowledge in SLA and bilingualism* (pp. 35–47). Washington, DC: Georgetown University Press.

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

Ellis, R. (1991). Grammaticality judgments and second language acquisition. *Studies in Second Language Acquisition*, *13*, 161–186.
Ellis, R. (1994). A theory of instructed second language acquisition. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 79–114). San Diego, CA: Academic Press.
Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language Learning*, *54*, 227–275.
Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172.
Ellis, R. (2009). Measuring implicit and explicit knowledge of a second language. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 31–64). 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*, 119–126.
Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.
Gass, S. (1983). The development of L2 intuitions. *TESOL Quarterly*, *17*, 273–291.
Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Mahwah, NJ: Erlbaum.
Goss, N., Ying-Hua, Z., & Lantolf, J. (1994). Two heads may be better than one: Mental activity in second-language grammaticality judgments. In E. Tarone, S. M. Gass, & A. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 263–286). Mahwah, NJ: Erlbaum.
Han, Y. (2000). Grammaticality judgment tests: How reliable and valid are they? *Applied Language Learning*, *11*, 177–204.
Han, Y., & Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research*, *2*, 1–23.
Hedgcock, J. (1993). Well-formed vs. ill-formed strings in L2 metalingual tasks: Specifying features of grammaticality judgments. *Second Language Research*, *9*, 1–21.
Hulstjin, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning. *Studies in Second Language Acquisition*, *27*, 129–140.
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*, 101–118.
Juffs, A. (2001). Psycholinguistically oriented second language research. *Annual Review of Applied Linguistics*, *21*, 207–222.
Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151.
Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.
Klein, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge.
Loewen, S. (2009). Grammaticality judgment tests and the measurement of implicit and explicit L2 knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 94–112). Bristol, UK: Multilingual Matters.
Loewen, S., & Erlam, R. (2006). Corrective feedback in the chatroom: An experimental study. *Computer Assisted Language Learning*, *19*, 1–14.
MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84–99.
Mandell, P. (1999). On the reliability of grammaticality judgment tests in second language acquisition research. *Second Language Research*, *15*, 73–99.
Martens, M. J. (1988). *Recognition and production of pronouns by francophone learners of English as a second language* (Unpublished master's thesis). Concordia University, Montreal.
Montrul, S. (2005). On knowledge and development of unaccusativity in Spanish L2. *Linguistics*, *43*, 1153–1190.
Montrul, S. (2011). Morphological errors in Spanish second language learners and heritage speakers. *Studies in Second Language Acquisition*, *33*, 163–192.
Mulaik, S. A. (2009). *Foundations of factor analysis* (2nd ed.) Boca Raton, FL: CRC Press.

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

Pennell, R. (1968). The influence of communality and *N* on the sampling distributions of factor loadings. *Psychometrika*, *33*, 423–439.
Sorace, A. (1985). Metalinguistic knowledge and language use in acquisition-poor environments. *Applied Linguistics*, *6*, 239–254.
Toth, P. D. (2006). Processing instruction and a role for output in second language acquisition. *Language Learning*, *56*, 319–385.
Tremblay, A. (2005). Theoretical and methodological perspectives on the use of grammaticality judgment tasks in linguistic theory. *Second Language Studies*, *24*, 129–167.
Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231–251.
Velicer, W. F., Peacock, A. C., & Jackson, D. N. (1982). A comparison of component and factor patterns: A Monte Carlo approach. *Multivariate Behavioral Research*, *17*, 371–388.
Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks, CA: Sage.

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press

APPENDIX

List of Structures in the GJTs

1. Determiner-noun agreement
2. Subject-verb agreement in present indicative
3. Subject-verb agreement in preterit
4. Subject-verb agreement in simple future
5. Stem-changing verbs in present indicative
6. Irregular verbs in preterit
7. *Ser* or *estar* with location
8. *Ser* or *estar* with profession
9. *Ser* or *estar* with place of origin
10. *Ser* or *estar* with personality traits
11. *Ser* or *estar* with mood or conditions
12. Imperfect with ongoing actions
13. Imperfect with habitual actions
14. Preterit with completed actions
15. Present subjunctive with noun clauses
16. Present subjunctive with adverbial clauses

https://doi.org/10.1017/S0272263113000041 Published online by Cambridge University Press