System 72 (2018) 215–225

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system

Exploring second language learners' grammaticality judgment performance in relation to task design features

Li-Ju Shiu [a, *], Şebnem Yalçın [c], Nina Spada [b]

[a] National Chi Nan University, 1 University Rd. Puli, 545 Nantou County, Taiwan
[b] University of Toronto, 252 Bloor St W, Toronto, ON M5S 1V6, Canada
[c] Boğaziçi University Faculty of Education, Department of Foreign Language Education, Bebek 34342 İstanbul, Turkey

ARTICLE INFO

Article history:
Received 12 January 2017
Received in revised form 14 December 2017
Accepted 14 December 2017
Available online 12 January 2018

Keywords:
Grammaticality judgment tasks
Task modality
Target structure difficulty
Task design features

ABSTRACT

This paper reports on an investigation of how second language (L2) learners' grammaticality judgment task (GJT) performance varies according to time constraints, task modality, and task stimulus in relation to two target features. One hundred and twenty EFL students were asked to judge items as grammatical or ungrammatical on four computer-based GJTs – two differing along the timed/untimed dimension and two differing along the aural/written dimension. Each GJT consists of 60 items (30 grammatical and 30 ungrammatical) focusing on two grammatical features in English, the passive voice and the past progressive, which were hypothesized to differ in terms of their learning difficulty. The results indicated that time constraints, task modality and task stimulus played a significant role in affecting L2 learners' GJT performance. Furthermore, although the learners performed better on the past progressive items, their GJT performance indicated similar patterns in relation to task design features across both target structures.

1. Introduction

Grammaticality judgment tests (GJTs)[1] have long been used to provide information about second language (L2) learning, including, for example, the investigation of adult L2 learners' access to Universal Grammar (e.g., Bley-Vroman, Felix, & Ioup, 1988), the critical period hypothesis in second language acquisition (SLA) (e.g., Johnson & Newport, 1989; Johnson, 1992), and L2 learners' use of different types of L2 knowledge (Bialystok, 1979; Bowles, 2011; Ellis & Loewen, 2007; Ellis, 2005; Godfroid, Loewen, Jung, Park, & Gass, 2015; Han & Ellis, 1998; Kim & Nam, 2016; Vafaee, Suzuki, & Kachisnke, 2017; Zhang, 2015). The extensive use of GJTs in SLA research derives from the hypothesis that they serve as promising measures of learners underlying linguistic competence. As a result, learners' GJT performance has been used to argue for and against different theoretical positions and empirical findings (e.g., Birdsong, 1989; Ellis, 1991; Han & Ellis, 1998; Hedgcock, 1993). The popularity of using GJTs in SLA research is also in part due to the fact that they are comparatively easy to administer to a large number of participants and they can assess knowledge of target features that are difficult to elicit in learners' production (Loewen, 2009).

* Corresponding author. 353 Wunsin First Rd. Nantun District, Taichung City 408, Taiwan.
E-mail address: ljshiu@ncnu.edu.tw (L.-J. Shiu).

[1] GJTs have been referred as "acceptability judgment tasks" in some studies because "acceptability judgment" was considered to be a more appropriate term to refer to learners' task performance (Ionin & Zyzik, 2014).

https://doi.org/10.1016/j.system.2017.12.004
0346-251X/

GJTs have different forms, including asking test takers to make grammaticality judgments, and/or to identify, correct and/or explain erroneous forms (Chaudron, 1983; Loewen, 2009). Among these task requirements, those asking L2 learners to judge the overall grammaticality of sentences have received more attention, particularly in recent studies exploring learners' use of different types of L2 knowledge (e.g., Ellis, 2009; Loewen, 2009; Godfroid et al., 2015; Gutierrez, 2013; Kim & Nam, 2016; Vafaee et al., 2017). It has been generally observed that L2 learners' performance on GJTs varies with learner-related factors (e.g., L2 proficiency level), linguistic features (e.g., target structures), and GJT task design variables (e.g., time constraints, task stimulus, and modality) (Bialystok, 1979; Ellis, 1991; Hedgcock, 1993; Godfroid et al., 2015; Gutierrez, 2013; Loewen, 2009; Murphy, 1997). Among the task design features, modality (i.e. visual versus aural) and target feature (i.e. more/less difficult to learn) have received less attention.

Due to the popularity of GJTs as research instruments to measure learners' L2 knowledge in the field of SLA (Loewen, 2009), it is essential to continue to explore how design features contribute to leaners' GJT performance. The current study builds on previous GJT research, exploring whether, and if so, to what extent the four variables—time constraints, task stimulus, task modality, and target features—affect L2 learners' grammaticality judgments.

2. Literature review

2.1. GJT design features and L2 learners' GJT performance

Below we review some of the GJT research indicating that learners' performance may vary with time constraints, task stimulus, task modality, and target features.

2.1.1. Time constraints

GJT research has generally found that L2 learners perform better on untimed GJTs than timed ones (Bowles, 2011; Ellis, 2005; Godfroid et al., 2015; Gutierrez, 2013; Han & Ellis, 1998; Han, 2000; Loewen, 2009; Mandell, 1999; Zhang, 2015). One often-given explanation for this finding is that with unlimited time, L2 learners, especially those receiving extensive amounts of classroom instruction, can take advantage of their L2 explicit knowledge while making judgments (e.g., Ellis, 2009; Loewen, 2009).

Ellis (2004) has proposed that when making grammaticality judgments, leaners are likely to go through three steps. First, they have to process semantically to understand the sentence (semantic processing). Then they need to detect if there is anything ungrammatical (noticing). If there is no grammatical error, they can make their judgment at this point. However, if learners notice something ungrammatical, they may reflect upon what (or maybe why) is not correct to confirm their initial detection of the ungrammatical element (reflecting). If learners are given enough time, they may go through the three steps before making a judgment. Accordingly, their GJT performance is better given the greater time allowance.

2.1.2. Task stimulus

Research using GJTs has also reported that L2 learners differ when judging grammatical and ungrammatical sentences. The majority of the studies employing GJTs have found that L2 learners perform better on grammatical rather than ungrammatical items (e.g., Bialystok, 1979, 1986; Gutierrez, 2013; Murphy, 1997; Kim & Nam, 2016; Loewen, 2009; Vafaee et al., 2017). However, a small number of studies using GJTs (e.g., Bley-Vroman et al., 1988; Gass, 1983) have found the opposite.

Reviewing a number of GJT studies, Hedgcock (1993) noted several possible factors that might affect learners' performance in judging grammatical and ungrammatical sentences, including, for example, the syntactic or semantic complexity of the test sentences, and influence of learning experience. To illustrate, errors may be easier to identify in sentences with simple structures than in sentences with complex structures because in the former, the errors may be more salient. Errors that occur in sentences with more complex semantic meanings might be more likely to be unnoticed because learners' attention might be more focused on the *meaning* rather than the *form* of the sentences. Erroneous forms of the features that have been extensively practiced might also be easier to detect than those that have been less frequently practiced because the former might have been overtly corrected.

Other views as to what factors might contribute to GJT judgments include Birdsong (1989) who argues that learners might tend to reject a grammatical sentence when unsure about its grammaticality. Ellis (1991) proposed that learners might consider an ungrammatical sentence to be grammatical due to lack of sufficient L2 knowledge. Gutierrez (2013) argued that L2 learners might resort to different types of L2 knowledge to respond to grammatical versus ungrammatical items; they might use explicit knowledge to respond to ungrammatical items and implicit knowledge in response to grammatical items.

2.1.3. Task modality

Researchers (e.g., Johnson, 1992; Penney, 1989; Wong, 2001) have assumed that task modality plays a role in influencing learners' GJT performance. In a comprehensive review of psychological research on modality differences, Penney (1989) argued that aural and visual verbal materials are processed in different parts of the memory system and by different mechanisms. McDonald (2000) also argued that decoding phonological stimuli is more demanding than decoding written stimuli because the former imposes more of a processing load. Wong (2001) explored whether modality affected how learners processed linguistic input, finding that learners had difficulty simultaneously attending to both form and meaning when the input was presented in an aural mode but not in a written mode.

Despite the finding regarding differences in processing auditory and visual stimuli, overall, there is insufficient research exploring the effect of task modality (aural vs. written) on task performance (Loewen, 2009; Rebuschat, 2013). Rebuschat (2013) expressed concern about the lack of studies investigating the effect of task modality, noting that "given that auditory input is central to both naturalistic and classroom L2 acquisition, it would be important to determine initially how learners acquire language implicitly from spoken language input. … [A] comparison of the role of listening and reading in implicit L2 learning would be worth pursuing" (p. 598).

Written GJTs have predominated L2 research with very few studies specifically investigating how modality affects learners' GJT performance (Marsden, Plonsky, Crowther, Gass, & Spinner, 2016). Among them, Johnson (1992) used both written and aural GJTs to investigate critical period effects with L2 leaners, finding that the learners performed significantly better on a written versus an oral GJT. Similar findings were reported in (Haig, 1991, cited in Murphy, 1997) that used both aural and written GJTs to explore leaners' L2 knowledge.

Murphy (1997) examined how modality would affect learners' performance on GJTs by analyzing their reaction time and accuracy scores. Four language groups participated in her study: English L2, French L2, English L1, and French L1. She found that learners performed better on the written GJT than the aural GJT, and that they made judgments faster on the written GJT than the aural GJT. The influence of modality was most obvious for the L2 learners, leading Murphy to argue that the learners' poorer performance on the aural GJT might be attributed to the higher demands of processing aural input than written input in an L2.

2.1.4. Target features

Some researchers (e.g., Bialystok, 1979; Ellis, 2004; Mandell, 1999) have assumed that variation in L2 learners' GJT performance may be attributable to the variability in the types of target structures tested. However, to date, little research has explored this question. Most studies employing GJTs focused on only one target feature (Marsden et al., 2016)[2] and while several recent GJT studies (e.g., Bowles, 2011; Loewen, 2009; Gutierrez, 2013; Kim & Nam, 2016; Vafaee et al., 2017) included multiple target structures, none specifically investigated whether and how different target features affect learners' GJT performance.

One exception is an early study by Bialystok (1979) who investigated L2 learners' GJT performance in relation to target features and time constraints. Bialystok administered two aural GJTs, which were identical except for their time limit: one asked the learners to respond within 3 s, whereas the other allowed 15 s to respond. The GJTs consisted of 24 French sentences (6 grammatical and 18 ungrammatical). The 18 ungrammatical sentences contained errors in three French form classes: adjectives, object pronouns, and verbs. Each form class consisted of three governing grammar rules, with a total of 9 governing rules. Two ungrammatical sentences targeted one governing rule. The three governing grammar rules in each form class were classified into "easy," "middle," and "difficult." For example, the rule applied only for adjectives (i.e., adjectives come before nouns) was categorized as "easy." The rule for a specific domain of adjectives (e.g., color adjectives follow nouns) was categorized as "middle." The rule that specified the gender agreement for adjectives was categorized as "difficult." The classification of the difficulty levels was based on Bialystok's own and native speaker participants' subjective judgments. The results indicated that the learners' GJT performance did not vary in relation to the target features on the 3-s GJT, but it did on the 15-s GJT. On the 15-s GJT, the learners performed better on the items targeting "easy" target structures than those targeting "difficult" target structures.

Despite evidence to suggest that target feature plays a role in GJT performance and ongoing speculation about this in the literature (e.g., Ellis, 2009; Han & Ellis, 1998), it has not yet been systematically investigated. That is, no study to our knowledge has compared whether L2 learners perform differently on GJTs targeting two target features hypothesized to differ in terms of difficulty.

2.2. Interaction effects of different design features

In addition to investigating the main effects of particular GJT design features on learner performance, researchers have also explored potential interaction effects (e.g., Bialystok, 1979; Loewen, 2009). For example, as discussed above, Bialystok's (1979) study showed interaction effects between time constraints and the difficulty level of target features. Loewen (2009) found an interaction effect of task stimulus and time constraints in his study where L2 learners of English performed significantly better on the grammatical items compared with ungrammatical items on timed GJTs, but similarly on both on the untimed GJTs.

It is also important to note that learners' GJT performance may vary with other factors such as their proficiency level (Gass, 1983), length of exposure to target languages (Garcia-Mayo, 2003), and working memory capacity (McDonald, 2008). However, these factors are beyond the scope of the current study and will not be addressed here. This review of the GJT research indicates that learners' GJT performance is associated with task design features and that most of this research has focused on task stimulus (grammatical/ungrammatical) and time constraints (timed/untimed). Less research has explored how learners' GJT performance varies according to modality (aural/written) or target features. The current study is an attempt to build on previous GJT research by responding to some of these gaps through an exploration of the following questions:

[2] According to Marsden et al. (2016), 66% of the GJT studies employed one target feature.

1. What effects do time constraints, task stimulus, and task modality have on L2 learners' GJT performance?
2. Does L2 learners' GJT performance vary depending on the nature of the target features?

3. Method

3.1. Participants and language proficiency measure

We recruited 181 English-as-a-foreign-language (EFL) learners from one university in Taiwan. The learners were from six intact EFL classes and 120 of them (76 female, 44 male) completed all the GJTs.[3] All participants spoke Mandarin Chinese as their mother tongue and the average age was 19.53 ($SD = 2.12$) (age range 18–34). There were also 54 native English-speaking participants (19 male; 35 female) who participated, all of whom were undergraduate students at a Canadian university with an average age of 28 ranging between 20 and 61. The majority of the L2 participants and the native English-speaking participants had some knowledge of at least one more language other than their mother tongue, but we do not have any information about the extent of their knowledge of other languages.

The Oxford Placement Test (OPT) was used to explore the L2 learners' general English proficiency. The OPT is a computer adaptive test, which consists of two sections—use of English and listening comprehension. The "use of English" section consists of 30 questions. This section involves multiple-choice questions and cloze tests, which test learners' knowledge of vocabulary, grammar, and reading comprehension. The listening comprehension section has 15 questions. In the listening section, the learners are asked to respond to short conversations, long conversations, and monologues. They are given the opportunity to listen to the questions twice. The learners' OPT responses are automatically calculated when they finish the test. The maximum score for each of the two sections is 120. The scores are reported corresponding to the Common European Framework of Reference (CEFR). The OPT does not set a time limit for the test-takers; the test-takers are allowed to complete the test in their own time.

Of the 120 EFL participants who completed all the GJTs, only 85 took the OPT. Results show that the mean score for the "use of English" section was 69.14 ($SD = 24.86$) and the mean score for the listening section was 50.01 ($SD = 20.91$). The OPT scores between 41 and 60 are interpreted as representing the B1 level of CEFR, and the scores between 61 and 80 suggest the B2 level. Accordingly, the average proficiency level of the Chinese EFL participants with respect to their listening ability was at the B1 level, whereas their ability of use of English was at the B2 level. The B1 and B2 levels refer to intermediate and upper intermediate levels respectively.

3.2. Target structures

The study targeted two English structures—the *be*-passive voice construction and the past progressive tense that were hypothesized to differ in terms of their learning difficulty. These two language structures were selected because they were the focus of investigation in other instructed SLA studies in our research project. L2 structure difficulty has been defined in different ways (e.g., linguistically, psycholinguistically, and pedagogically) (Housen & Simonens, 2016) and there is no universally-accepted definition. For the purpose of this study, target structure difficulty was defined in terms of structure formation, input frequency, and phonological saliency. Following Hulstijn and de Graaff (1994), the passive was considered to be the more difficult structure because its production involves a higher number of grammatical operations compared to the past progressive. Its lack of frequency in the input[4] (Hinkel, 2004) and the fact that it is acquired relatively late with L1 learners (Kirby, 2010) are additional reasons contributing to its difficulty. The past progressive was characterized as the less difficult structure because of its transparent form-meaning relationship and greater frequency in the input (Collins, Trofimovich, White, Cardoso, & Horst, 2009; Révész, 2009). The past progressive is also more salient as it is realized by a free morpheme (was/were) and syllabic bound morpheme (*-ing*) (Révész, 2009). In addition, there are no allomorphs for the present participle *–ing* marker that would reduce its transparency. In contrast, the variety of allomorphs of the past participle in the formation of the passive renders it a more difficult structure. For more discussion of the characterization of these features as more or less difficult to learn, see Yalcin and Spada (2016).

3.3. Research instruments

Four GJTs were administered. All of the GJTs have two versions, differing in the order of item presentation to avoid a potential task effect. What follows is a detailed description of all the tests.

3.3.1. Timed aural grammaticality judgment test

The timed aural GJT (AGJT) consists of 60 items, with 24 targeting the passive construction, 24 targeting the past progressive, and 12 distractors targeting other grammatical features. The passive items vary in terms of length (10-14 syllables,

[3] The number of the participants who took the tests slightly varied. One hundred and fifty-five learners took the timed aural GJT, 151 took the timed written GJT, 157 took the untimed aural GJT, and 177 took the untimed written GJT.

[4] Nonetheless, the English passive is used more frequently than the Chinese (the participants' L1) passive (McEnery & Xiao, 2005).

with an average of 11.96 syllables), accuracy (12 grammatical and 12 ungrammatical), and tense (8 present, 8 past, 8 present perfect). The passive items are all simple sentences. The passive items include 12 regular verbs and 12 irregular verbs. The ungrammatical items focus on two types of errors: omitting auxiliary verb *be* (e.g., *Every year, many children reported missing.*), and using the bare form of the verb instead of past participle (e.g., *The taxi has been park at the airport for three months.*). With reference to verb types (regular vs. irregular), the error types of the passive items can be divided into four categories abbreviated as: (a) regular *be*, (b) regular participle, (c) irregular *be*, and (d) irregular participle.

The 24 past progressive items are also evenly divided between grammatical and ungrammatical sentences. In order to address differences in lexical aspect (Vendler, 1967), 12 items included verbs of accomplishment and 12 included verbs of activity. The length of the past progressive items ranged between 12 and 16 syllables, with an average length of 13.46 syllables. Twelve items are grammatical, while the other 12 are ungrammatical items, targeting two error types: (1) missing auxiliary (e.g. *While the girl sitting outside, it started raining*), and (2) present auxiliary (e.g., *She is reading a book at 4 yesterday afternoon*). Sixteen of the past progressive items consist of subordinate clauses that indicate the action taking place at a certain time in the past (e.g., *When I met my husband, I was traveling in France.*), whereas the rest 8 sentences are simple sentences.[5] The differences between the two target features are taken into consideration in the analysis of the data discussed below.

A vocabulary frequency analysis of the GJT items indicated that 91.88 of the words appear in the first two frequency bands (i.e., "K1 and K2 words," 83.37% and 8.51%, respectively) of Nation's word lists (Laufer & Nation, 1995). "K1 words" refer to the most frequent 1000 English words, and "K2 words" refer to the second most frequent 1000 English words (Cobb, 2013). Thirty-nine words (7.54%) of the words are off-list words. Most of the off-list words are proper nouns (e.g., Japan, John). The remaining words (e.g., supermarket, beer) are considered to be known by the participants as these words are covered in the high school English textbooks.[6]

Following Loewen (2009), we used the response time of native speakers (NS) of English as a baseline to calculate the time limit for each aural GJT item. Twenty-seven NS participants (10 male, 17 female; aged 28.81, ranged from 20 to 61, $SD = 9.73$) completed the AGJT[7] and their response times were recorded. Concerned that the NS participants might respond before they heard the sentence in its entirety, they were asked to respond after a beep that sounded immediately after the recorded sentence was finished. The beep sound took 0.25 s. Like Ellis (2005), Loewen (2009), and Vafaee et al. (2017), for each individual item, we calculated the median response time to avoid outlier effects, and then an additional 20% of the time it took NS to respond was added to the time given to L2 learners to respond. Accordingly, the time limit set for each timed AGJT item was the sentence reading time, the time for the beep sound plus the obtained response time. Thus, the amount of time that L2 learners were given to hear and respond to each item ranged from 4.08 s for Item 24 (*I grew up in London.*)[8] to 8.04 s for Item 11 (*Tom heard about the plane crash while he listening to the news last night.*).

The timed aural GJT was delivered online. After the participants clicked on the link provided, they saw a "welcome" page, which was followed by a page asking for some background information (including ID number, gender, L1, age, length of English learning). The next page included the directions stated as follows: "This is a timed test. You will only have a few seconds to respond. Each sentence will be heard only once. Please make your choice as quickly as possible." Before the start of the actual test, the participants were given four practice items to familiarize themselves with the speeded nature of the test. When the actual test started, for each item, the participants saw a page with the item number, below which were three options "Correct," "Incorrect," and "Not sure." Once the participants clicked on the option that they chose, the next page would appear. If they did not respond within the time limit set for the item in question, the next page would automatically appear, and their answer was classified as "no response."

3.3.2. Timed written grammaticality judgment test

The timed written GJT (WGJT) is virtually identical to the timed aural GJT except it was delivered in the written mode. Another 27 NS participants[9] (9 male, 18 female; average age 28, ranged from 20 to 58, $SD = 10.25$) completed the WGJT. The median number of seconds that it took the NS participants to read and respond was calculated. An additional 20% of the median time was given to the time for L2 learners to read and respond. The amount of time that the learners were given to read and respond to the items ranged from 2.16 s for Item 24 (*I grew up in London*) to 7.44 s for Item 22 (*In Los Angeles, many new cars were stolen last year*).

The written GJT was also delivered online. The information page and the directions page were the same as those for the timed AGJT. The participants were also provided with four practice items to familiarize themselves with the speeded nature of the test. After the test started, for each item, the participants saw a page with the item number, the item, and three options

[5] Although it was desirable to balance the passive and the past progressive items in terms of their length, error types and sentence pattern (simple vs. complex structure), it was difficult to have a strict control over these variables.

[6] The inclusion of the words in the high school textbooks was verified with two Taiwanese high school EFL teachers.

[7] The mean accuracy percentage of the NS participants on the AGJT was 91.00 ($SD = 4.77$). NS participants performed significantly better on the past progressive items (mean accuracy percentage $M = 93.00$, $SD = 4.81$) than on the passive items ($M = 89.00$, $SD = 8.02$), $t(26) = -2.19$, $p < .05$.

[8] This is a distractor item.

[9] These 27 participants were different from those who did the aural GJT. The mean accuracy percentage of these NS participants on the written GJT was 92.15 ($SD = 5.76$). NS participants performed similarly well on the passive items ($M = 92.41$, SD = 6.23) and the past progressive items ($M = 92.07$, $SD = 6.79$).

"Correct," "Incorrect," and "Not Sure." The next page appeared after the participants clicked on the option they chose or if they did not respond within time limit.

3.3.3. Untimed aural grammaticality judgment test

The untimed aural GJT was the same as the timed aural GJT except that there were no time constraints for learners' responses. The participants could take their time to respond and to listen to the item repeatedly if they felt necessary before responding. Because in the untimed written GJT, the participants were able to read a sentence more than once, to make the task demands of both untimed GJTs more parallel, repetitive listening was also allowed in the untimed aural GJT. The frequency of repeatedly listening to the sentence was recorded. The directions for the untimed AGJT were "After you hear the sentence, please choose 'Correct,' 'Incorrect, or 'Not Sure.' If you would like to hear the sentence again, press 'Listen Again.' You can take as much time as you need to make your decision." After the learner responded, the next question automatically appeared.

3.3.4. Untimed written grammaticality judgment test

The untimed written GJT is the same as the timed WGJT except that there are no time constraints for learners' responses. The directions for the untimed WGJT were "You can take as much time as you need to make your choice."

3.4. Data collection procedures

The four GJTs were piloted on 15 EFL learners before they were administered to the participants of the current study. These 15 EFL learners were recruited from the same university. No items were replaced after the pilot. The timed AGJT was administered first followed by the timed WGJT. There was a 30-min interval between the administrations of the two tests. One week after the participants completed the timed GJTs, they completed the untimed AGJT followed by the untimed WGJT. There was also a 30-min interval between the administrations of the two tests. The AGJT was administered before the WGJT because it was assumed that the aural stimuli were more transitory than the written stimuli. Therefore, administering the AGJT before the WGJT would decrease the possibility of memory effect. All tests were administered during regular class hours.[10]

3.5. Data analyses

The four GJTs were scored in terms of accuracy, with 1 point for a correct response and 0 point for incorrect and no response. The maximum score for each GJT was 48. The option "Not sure" was considered to be incorrect.[11] "No response" items accounted for 13% and 18% of all the responses to the timed AGJT and timed WGJT respectively.[12]

The reliability of the four GJTs was calculated based on the 120 EFL students' data, using Cronbach's alpha. The reliability coefficients of the timed AGJT, timed WGJT, untimed AGJT, and untimed WGJT were 0.80, 0.87, 0.81, and 0.86, respectively. Descriptive statistics of the EFL participants were calculated for the four GJTs. Bivariate correlations were computed to examine the inter-correlations among the four tests and the Oxford Placement Test (including the listening section, the "use of English" section, and the combination of the two sections). Bivariate correlations were also computed to examine the relationships among the grammatical and ungrammatical items of the four GJTs. Repeated-measures ANOVA tests were performed on the 120 EFL learner data. Given that the items of the two target features are not identical in terms of their length, error types and sentence pattern (i.e., simple versus complex), the bivariate correlations and the repeated-measures ANOVA tests were conducted separately for the passive structure and the past progressive structure. The participants' GJT performance was also examined in relation to the different error types included in the ungrammatical items of the two target features.

4. Results

Table 1 presents the descriptive statistics for all the tests for the 120 EFL participants. As Table 1 indicates, overall the learners performed the best on the untimed WGJT ($M = 39.02$, $SD = 6.09$) and the least well on the timed AGJT ($M = 23.90$, $SD = 5.69$). In terms of task stimulus, the learners generally performed better on the grammatical items than on the ungrammatical items. With regard to the learners' performance on the passive structure and the past progressive, it was

[10] Because the participants in the current study were from intact EFL classes, the instructors suggested administering the tests in their regular class hours, and the participants agreed.

[11] The option "Not sure" was included with a consideration that the learners might not be 100% sure about their choice. This option is considered to be an incorrect response because of an assumption that the learners might not have sufficient L2 knowledge of the target features. However, the "Not sure" option accounted for very low percentage of all the responses: 5%, 3%, 07%, and 0.3% respectively to the timed AGJT, the untimed AGJT, the timed WGJT, and the untimed WGJT.

[12] Blank items are mainly from the timed GJTs. For each item of the timed GJTs, if the participants did not respond within the time limit, the next item would automatically appear. Thus, the item that was not responded to became a blank item. The untimed GJTs did not have blank items because the next item did not appear until the participants responded to the previous item.

Table 1
Descriptive statistics.

	EFL ($N = 120$)					
	Total		Passive		Past Progressive	
	M[a]	SD	M	SD	M	SD
Timed AGJT	23.90	5.69	10.92	3.02	12.98	3.60
Grammatical	16.69	3.98	8.20	2.33	8.49	2.24
Ungrammatical	7.21	3.46	2.72	2.10	4.49	2.35
Timed WGJT	27.10	7.08	12.45	3.79	14.66	3.89
Grammatical	18.51	3.79	9.05	2.27	9.46	2.03
Ungrammatical	8.60	4.82	3.43	2.73	5.20	2.81
Untimed AGJT	32.27	5.79	13.94	2.81	18.33	3.88
Grammatica	19.74	3.16	9.63	1.73	10.13	1.98
Ungrammatical	12.53	4.58	4.32	2.55	8.21	2.81
Untimed WGJT	39.02	6.09	18.11	3.79	20.92	3.17
Grammatical	21.51	2.83	10.71	1.50	10.80	1.74
Ungrammatical	17.52	4.68	7.40	3.17	10.12	2.34

[a] The maximum score for each GJT was 48. The maximum score for each target feature is 24, and the maximum score for the grammatical/ungrammatical items of GJT was 12.

observed that the learners did the most poorly on the ungrammatical items for the passive structure on the timed AGJT ($M = 2.72$, $SD = 2.10$), and they did the best on the grammatical items for the past progressive structure on the untimed WGJT ($M = 10.80$, $SD = 1.74$). A look at Table 1 indicates that the differences between the grammatical items and the ungrammatical items on the aural GJTs for the passive structure were all larger than those on the aural GJTs for the past progressive structure.

Table 2 shows Pearson correlation coefficients for the learners' performance on the GJTs, and Oxford Placement Test.[13] Because only 85 out of 120 Chinese participants completed the OPT, the Pearson correlation coefficients were computed only with the data from this group. As indicated, among the four GJTs, the untimed AGJT has the highest correlation with the OPT total scores ($r = 0.68$). Likewise, for the listening section and the "use of English" section of the OPT, the untimed AGJT has comparatively higher correlations with the listening section ($r = 0.62$) and the "use of English" section ($r = 0.59$). As shown in Table 2, most of the four GJTs are moderately correlated with each other. The highest correlation was between untimed AGJT and untimed WGJT ($r = 0.86$). The lowest correlation was between the timed AGJT and untimed WGJT ($r = 0.48$).

Tables 3 and 4 present Pearson correlation coefficients for the learners' performance on the grammatical and ungrammatical items of the four GJTs, respectively for the passive structure items and the past progressive structure items. Overall, the correlations were small or medium. Across all the tests, the grammatical-grammatical item correlations and the ungrammatical-ungrammatical item correlations were slightly higher than the grammatical-ungrammatical item correlations.

Table 5 shows the results of the repeated measures ANOVA for all the target items. As Table 5 shows, there was a significant main effect for time constraint, $F(1, 119) = 616.45$, $p < .05$, $\eta_p^2 = 0.84$, for modality, $F(1, 119) = 156.64$, $p < .05$, $\eta_p^2 = 0.57$, and for task stimulus, $F(1, 119) = 575.04$, $p < .05$, $\eta_p^2 = 0.83$. These results suggest that the participants performed better on (a) the untimed GJTs than the timed GJTs, (b) the written GJTs than the aural GJTs, and (c) the grammatical items than the ungrammatical items. There was also a significant interaction between time constraint and modality, $F(1, 119) = 29.12$, $p < .05$, $\eta_p^2 = 0.20$, for time constraint and stimulus, $F(1, 119) = 76.97$, $p < .05$, $\eta_p^2 = 0.40$, and for modality and stimulus, $F(1, 119) = 18.90$, $p < .05$, $\eta_p^2 = 0.14$. There was also an interaction for time constraint, modality and stimulus, $F(1, 119) = 33.37$, $p < .05$, $\eta_p^2 = 0.22$.

Table 6 shows the results of the repeated measures ANOVA separately for the passive structure and the past progressive. As Table 6 shows, there was a significant main effect of time constraint for the passive structure, $F(1, 119) = 369.79$, $p < .05$, $\eta_p^2 = 0.76$, and for the past progressive structure, $F(1, 119) = 458.13$, $p < .05$, $\eta_p^2 = 0.79$. There was a significant main effect of modality for the passive, $F(1, 119) = 92.98$, $p < .05$, $\eta_p^2 = 0.44$, and for the past progressive, $F(1, 119) = 112.12$, $p < .05$, $\eta_p^2 = 0.49$. There was also a significant main effect of task stimulus for the passive, $F(1, 119) = 569.85$, $p < .05$, $\eta_p^2 = 0.83$, and for the past progressive, $F(1, 119) = 244.18$, $p < .05$, $\eta_p^2 = 0.67$. These results suggest that for both target structures, the participants performed better on (a) the written GJTs than the aural GJTs, (b) the untimed GJTs than the timed GJTs, and (c) the grammatical items than the ungrammatical items.

Results also revealed several significant interaction effects, including:

1. A significant interaction between time constraint and modality for the passive structure, $F(1, 119) = 44.92$, $p < .05$, $\eta_p^2 = 0.27$, and for the past progressive structure, $F(1, 119) = 4.90$, $p = .029$, $\eta_p^2 = 0.04$.

[13] One reviewer suggested that we report the corrected correlation coefficients for reliability. We took on the suggestion and calculated the adjusted correlation coefficients, using a formula of rxy / sqrt (rxx * ryy) (rxy = the raw correlations of measure x and y. rxx, ryy = reliability of measure x and measure y). These are indicated in Table 2 for all four GJTs. However, because the OPT was an online test that provided us with only the final total scores, it was not possible to calculate reliability. Therefore, in Table 2, only the raw correlation coefficients between the OPT and the four GJTs are indicated.

Table 2
Pearson correlations for GJTs and OPT (N = 85).

	1	2	3	4	5	6	7
1. Timed AGJT	1						
2. Untimed AGJT	.73**	1					
3. Timed WGJT	.54**	.65**	1				
4. Untimed WGJT	.48**	.86**	.57**	1			
5. Oxford Listening	.40**	.62**	.40**	.45**	1		
6. Oxford Use of English	.38**	.59**	.38**	.50**	.57**	1	
7. Oxford Total	.44**	.68**	.44**	.53**	.86**	.91**	1

** $p < .01$.

Table 3
Correlations among the grammatical and ungrammatical items of the four GJTs for the Passive Items.

	1	2	3	4	5	6	7	8
1. Timed AGJT G	1.00							
2. Timed AGJT UG	-.07	1.00						
3. Untimed AGJT G	.31**	.07	1.00					
4. Untimed AGJT UG	.15	.33**	-.18*	1.00				
5. Timed WGJT G	.26*	-.09	.21*	.12	1.00			
6. Timed WGJT UG	-.02	.26**	-.08	.45**	.14	1.00		
7. Untimed WGJT G	.27*	-.06	.58**	-.04	.26**	.10	1.00	
8. Untimed WGJT UG	.07	.10	-.01	.52**	.26**	.57**	.22*	1.00

Note. AGJT = aural grammatical judgment test; WGJT = written grammatical judgment test; G = grammatical sentences; UG = ungrammatical sentences.
* $p < .05$, ** $p < .01$.

Table 4
Correlations among the grammatical and ungrammatical items of the four GJTs for the Past Progressive Items.

	1	2	3	4	5	6	7	8
1. Timed AGJT G	1.00							
2. Timed AGJT UG	.23*	1.00						
3. Untimed AGJT G	.37**	.22*	1.00					
4. Untimed AGJT UG	.43**	.45**	.27**	1.00				
5. Timed WGJT G	.46**	.25**	.29**	.31**	1.00			
6. Timed WGJT UG	.35**	.55**	.28**	.44**	.27**	1.00		
7. Untimed WGJT G	.40**	.13	.65**	.21**	.32**	.11	1.00	
8. Untimed WGJT UG	.20*	.31**	.12	.57**	.16	.30**	.18	1.00

Note. AGJT = aural grammatical judgment test; WGJT = written grammatical judgment test; G = grammatical sentences; UG = ungrammatical sentences.
* $p < .05$, ** $p < .01$.

Table 5
Repeated measures ANOVA on all the target items.

	F	*p*	η_p^2
Time Constraint	616.45	.00	.84
Modality	156.64	.00	.57
Stimulus	575.04	.00	.83
Time*Modality	29.12	.00	.20
Time*Stimulus	76.97	.00	.40
Modality*Stimulus	18.90	.00	.14
Modality*Time*Stimulus	33.37	.00	.22

$N = 120$, $df = 1$.

2. A significant interaction between modality and stimulus for the passive structure, $F(1, 119) = 16.52$, $p < .05$, $\eta_p^2 = 0.12$, and for the past progressive structure, $F(1, 119) = 6.15$, $p = .015$, $\eta_p^2 = 0.05$ for the past progressive structure.
3. A significant interaction between time constraint and stimulus for the passive structure, $F(1, 119) = 18.28$, $p = .002$, $\eta_p^2 = 0.13$, and for the past progressive structure, $F(1, 119) = 109.86$, $p = < .05$, $\eta_p^2 = 0.48$.
4. A significant interaction between modality, timing, and grammaticality for the passive structure, $F(1, 119) = 23.90$, $p < .05$, $\eta_p^2 = 0.17$, and for the past progressive structure, $F(1, 119) = 13.60$, $p < .05$, $\eta_p^2 = 0.10$.

Table 6
Repeated measures ANOVAs for the passive items and the past progressive items.

	Passive			Past Progressive		
	F	p	η_p^2	F	p	η_p^2
Time Constraint	369.79	.000	.76	458.13	.000	.79
Modality	92.98	.000	.44	112.12	.000	.49
Stimulus	569.85	.000	.83	244.18	.000	.67
Time*Modality	44.92	.000	.27	4.90	.029	.04
Time*Stimulus	18.28	.000	.13	109.86	.037	.48
Modality*Stimulus	16.52	.000	.12	6.15	.015	.05
Time*Modality*Stimulus	23.90	.000	.17	13.60	.000	.10

$N = 120$, $df = 1$.

With regard to the effect sizes, both features have similar large effect sizes for the main effects. However, the interaction effect sizes are smaller. Some of the effect sizes of the interaction effects for the two features tend to be medium or approaching small.[14]

Comparisons of participants' performance on the passive items and the past progressive items, across all the four GJTs revealed that the learners performed significantly better on the past progressive items than the passive items. This suggests some support for the hypothesis that the passive was more difficult than the past progressive. However, because the error types of the two target features could be argued to be more or less difficult (e.g. focus on the auxiliary with the past progressive and focus on the main verb with the passive), the learners' GJT performance was examined in relation to the different error types in the ungrammatical items for both features. Results showed that across all four GJTs, there were no significant differences in the learners' performance with respect to the different error types for the passive and past progressive.

5. Discussion

The ANOVA results showed significant large main effects for modality, timing, and stimulus for both target features. There were also several significant interaction effects. The finding that learners performed better on the untimed GJTs than on the timed GJTs is in line with the results of previous studies (Bowles, 2011; Ellis, 2005; Gutierrez, 2013; Han & Ellis, 1998; Han, 2000; Loewen, 2009; Mandell, 1999; Zhang, 2015). The explanation for these findings is that when learners are given sufficient time to respond, this creates an opportunity for the learners to reflect upon the grammatical correctness of the sentences when making their judgment (Bialystok, 1979; Ellis, 2005; Loewen, 2009). This may be particularly true for learners who have had extensive amounts of grammar instruction as the learners in the current study. However, as Loewen (2009) has noted, we cannot rule out the possibility that the lower mean percentage on the timed GJTs was because of their "speeded nature." That is, the timed GJTs did not provide enough time for learners to access their knowledge regardless of the amount of knowledge they possessed about the target features.

Corresponding to the findings of many GJT studies (e.g., Bialystok, 1979, 1986; Gutierrez, 2013; Murphy, 1997; Kim & Nam, 2016; Loewen, 2009; Vafaee et al., 2017), our findings revealed that the learners performed significantly better on the grammatical than on the ungrammatical items of the GJTs. One possible explanation for this is that the learners did not have sufficient L2 knowledge of target features to detect grammatical errors in the sentences. Thus, they tended to consider the ungrammatical sentences to be grammatically correct. Such a "response bias" caused by insufficient L2 knowledge might have led to the less accurate performance on the ungrammatical items. As Birdsong (1989) has argued, any decision-making tasks may be influenced by "guesswork" or "response bias." Notwithstanding, to what extent "guesswork" or "response bias" may influence the students' decision-making is an empirical question that warrants further exploration.

The finding regarding task stimulus is contrary to that reported in Bley-Vroman et al. (1988) in which learners judged ungrammatical items more accurately than grammatical items. A possible reason for this discrepancy may be related to differences in the proficiency level of learners. The participants in Bley-Vroman et al. (1988) were advanced ELS learners living in the U.S., whereas the participants in the current study were intermediate EFL learners in Taiwan. Other variables (e.g., the modality of the tasks) may have also contributed to these differences.

The results also showed that the increase in the GJT scores (a) from timed to untimed GJTs and (b) from aural to written GJTs was greater when the items were ungrammatical compared to when the items were grammatical. A similar pattern has been reported in Gutierrez (2013) and Loewen (2009). This suggests that learners may be processing ungrammatical sentences differently from grammatical sentences (Gutierrez, 2013).

With regard to the effect of task modality, consistent with Murphy (1997), (Haig, 1991; cited in Murphy), and Johnson's (1992) findings, we found that the learners' performance was less accurate on the aural GJT than the written GJT. Additionally, like Murphy, we also found that there was a significant interaction effect for task stimulus and task modality; the

[14] We have used the standard rule of thumb for interpreting the magnitude of partial eta squared effect sizes (i.e., 0.01 small; 0.06 medium; 0.14 large) (Gray & Kinnear, 2012) in this study, but see Plonsky and Oswald (2014) for a discussion of the factors to consider when gauging the relative magnitude of effect sizes (specifically Cohen's d) in L2 research.

mean difference between the grammatical section and the ungrammatical section was largest on the timed AGJT, but smallest on the untimed WGJT. As Danks (1980) argues, learners have more control over processing sentences when items are presented visually than when they are presented aurally. Thus, learners might have more control over carefully processing items during WGJTs than during AGJTs.

Another possible explanation for the learners' poorer performance on the aural GJT might be due to their prior L2 learning experience. The learners in the study had received much more written than aural input. Therefore, they were assumed to have better reading than listening skills. Indeed, this assumption seems to be supported by the results of the Oxford Placement Test, showing that the learners did better on the "use of English" section than the listening section. Additionally, as indicated in Table 2 that shows the correlations between the learners' GJT performance and their OPT performance, the finding that the untimed AGJT had the highest correlation with the listening section of the OPT also somewhat supports the assumption that the learners' aural GJT performance is associated with their listening ability because the test condition for the listening section and the untimed AGJT was similar in the sense that both tests did not set time constraint for the learners to respond.

In the current study, we defined learning difficulty in terms of perceptual salience, frequency in the input, inherent structural complexity and early/later acquired. The hypothesized difference in the difficulty level of the two target features seemed partially supported by the finding that the learners performed significantly better on the past progressive items than on the passive items, despite the fact that the average length of the past progressive items was longer than that of the passive items. Because the items for the two language features were not identical in terms of sentence length and type, ANOVAs were conducted separately for the passive and past progressive. The results showed that the main and interaction effects were similar for both target structures. That is, both features had similar patterns in terms of the effects of time constraints, task modality and task stimulus. For both features, the learners performed better (a) on the untimed GJTs than the timed GJTs, (b) on the written GJTs than the aural GJTs, and (c) on the grammatical items than the ungrammatical items.

With regard to effect sizes, the interaction effects of Time*Modality and Modality*Stimulus for the past progressive are small ($\eta_p^2 = 0.04$ and 0.05, respectively). This may provide support for the hypothesis that the past progressive was an easier structure for the participants so that neither modality nor amount of time made a difference with respect to their GJT performance. However, this is speculative and more research is warranted to contribute to our understanding of whether/how learners' GJT performance varies in relation to the nature of target features. Moreover, even though a comparison of the more or less difficult error types in the ungrammatical items for both target features indicated no differences in learners' GJT performance, questions remain as to whether the errors capture the 'core' or 'intrinsic' differences in difficulty between the two structures.

6. Limitations and implications

The current study was conducted to explore how time constraints, task stimulus, task modality, and target features might affect L2 learners' GJT performance. The results showed that although the learners performed better on the past progressive items than the passive items, their GJT performance indicated similar patterns in relation to task design features (i.e., time constraint, modality, and stimulus) across both target structures. Additionally, time constraints, task modality and task stimulus all played a role in influencing learners' GJT performance.

There are several limitations to this study. First, the order in which the tests were administered was not counter balanced. All of the participants took the four GJTs in the same order: timed AGJT→ timed WGJT → untimed AGJT →untimed WGJT. The order was not counter-balanced due to a concern that the written stimuli might have generated a more powerful memory effect if given before the aural test with only a 30-min interval between the administrations of the two GJTs. Second, the sentence length, error types, and sentence patterns lacked strict control. Third, to keep the design parallel for both the untimed aural and written GJTs, a decision was made to allow for repeated listening in the untimed aural GJT. This may have introduced a confounding variable and could be a focus of investigation in future GJT research. Fourth, the participants of the study were EFL learners whose L1 was mainly Mandarin Chinese and there was a lack of variation in terms of the participants' English proficiency level (i.e., they were mainly at B1 and B2 level). Thus, the results cannot be generalized to other L2 populations. Furthermore, although the majority of learners had some knowledge of another language, the extent of this knowledge is unknown and thus how this might have contributed to differences in performance.

Nonetheless, the findings of the current study contribute further to our understanding of how GJT task design features affect L2 learners' performance. In particular, it is hoped that the focus on two under-researched variables – modality and target feature – will help to move the methodological research agenda forward given the popularity of the use of GJTs in L2 research.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

References

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*. Berlin, Germany: 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.
Chaudron, C. (1983). Research in metalinguistic judgment: A review of theory, methods and results. *Language Learning, 33*, 343–377.
Cobb, T. (2013). *Complete lexical tutor*, 2013 v.6.2. http://www.lextutor.ca.
Collins, L., Trofimovich, P., White, J., Cardoso, W., & Horst, M. (2009). Some input on the easy/difficult grammar question: An empirical study. *The Modern Language Journal, 93*, 336–353.
Danks, J. (1980). Comprehension in listening and reading: Same or different? In J. Danks, & K. Pezdek (Eds.), *Reading and understanding* (pp. 1–39). Newark, DE: International Reading Association.
Ellis, R. (1991). Grammaticality judgments and second language acquisition. *Studies in Second Language Acquisition, 13*, 161–186.
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). Tonawanda, NY: 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.
Garcia-Mayo, M. P. (2003). Age, length of exposure and grammaticality judgements in the acquisition of English as a foreign language. In M. P. Garcia-Mayo, & M. L. Garcia-Lecumberri (Eds.), *Age and the acquisition of of english as a foreign language* (pp. 94–114). Clevedon: Multilingual Matters.
Gass, S. (1983). The development of L2 intuitions. *Tesol Quarterly, 17*, 273–291.
Godfroid, A., Loewen, S., Jung, S., Park, J., & Gass, S. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: Evidence from eye-movement patterns. *Studies in Second Language Acquisition, 37*, 269–297.
Gray, C., & Kinnear, P. (2012). *IBM SPSS19 statistics made simple*. New York, NY: Psychology Press.
Gutierrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition, 35*, 423–449.
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.
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.
Hinkel, E. (2004). Tense, aspect and the passive voice in L1 and L2 academic texts. *Language Teaching Research, 8*, 5–29.
Housen, A., & Simoens, H. (2016). Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition, 38*, 163–175.
Hulstijn, J., & de Graaff, R. (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review, 11*, 97–112.
Ionin, T., & Zyzik, E. (2014). Judgment and interpretation tasks in second language research. *Annual Review of Applied Linguistics, 34*, 37–64.
Johnson, J. S. (1992). Critical period effects in second language acquisition: The effects of written versus auditory materials on the assessment of grammatical competence. *Language Learning, 42*, 217–248.
Johnson, J. S., & Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*, 60–99.
Kim, J., & Nam, H. (2016). Measures of implicit knowledge revisited: Processing modes, time pressure, and modality. *Studies in Second Language Acquisition*, 1–27.
Kirby, S. (2010). Passives in first language acquisition: What causes the delay? *University of Pennsylvania Working Papers in Linguistics, 16*, 109–117.
Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written productions. *Applied Linguistics, 16*, 307–322.
Loewen, S. (2009). Grammaticality judgment tests and the measurement of implicit and explicit L2 knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 94–112). Tonawanda, NY: Multilingual Matters.
Mandell, O. (1999). On the reliability of grammaticality judgment tests in second language acquisition research. *Second Language Research, 15*, 73–99.
Marsden, E., Plonsky, L., Crowther, D., Gass, S., & Spinner, P. (2016, September). *A methodological synthesis of judgment tasks in second language research. Paper presented at the Second Language Research Forum (SLRF), Teachers College*. New York: Columbia University.
McDonald, J. L. (2000). Grammaticality judgments in a second language: Influences of age of acquisition and native language. *Applied PsychoLinguistics, 21*, 395–423.
McDonald, J. L. (2008). Grammaticality judgments in children: The role of age, working memory, and phonological ability. *Journal of Child Language, 35*, 247–268.
McEnery, T., & Xiao, R. (2005). *Passive constructions in English and Chinese: A corpus-based contrastive study*. Retrieved January 20, 2010, from http://eprints.lancs.ac.uk/63/.
Murphy, V. (1997). The effect of modality on a grammaticality judgment task. *Second Language Research, 13*, 34–65.
Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition, 17*, 398–422.
Plonsky, L., & Oswald, F. L. (2014). How big is "Big"? Interpreting effect sizes in L2 Research: Effect sizes in L2 research. *Language Learning, 64*, 878–912. https://doi.org/10.1111/lang.12079.
Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning, 63*, 595–626.
Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition, 31*, 437–470.
Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). Validating grammaticality judgments: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, 59–95.
Vendler, Z. (1967). *Linguistics in philosophy*. Ithaca, NY: Cornell University Press.
Wong, W. (2001). Modality and attention to meaning and form in the input. *Studies in Second Language Acquisition, 23*, 345–368.
Yalcin, S., & Spada, N. (2016). Language apptitude and grammatical difficulty: An EFL classroom-based study. *Studies in Second Language Acquisition, 38*, 239–263.
Zhang, R. (2015). Measuring university-level L2 learners' implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition, 37*, 457–486.