THE JOURNAL OF ASIA TEFL
Vol. 17, No. 2, Summer 2020, 349-362
http://dx.doi.org/10.18823/asiatefl.2020.17.2.3.349

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511 © 2004 AsiaTEFL.org. All rights reserved.

Grammaticality Judgment Task: Reliability and Scope

Muhammad Asif Qureshi
Zayed University, Abu Dhabi, UAE

Although extensively used, concerns have been expressed about the reliability and generalizability of grammaticality judgment tasks (GJTs; Alanazi, 2015). It has been argued that learners might guess grammaticality based on feel (Bialystock, 1979) and that the GJT-based results ignore grammatical complexity (Ellis, 1991). While several studies have attempted to validate the results of a GJT against other tasks, the tasks used in these studies required language production (e.g., Leow, 1996; Mandell, 1999), which is different from error identification and correction. The current study used an editing task (ET) to validate learners' performance on a GJT. An editing task, like a GJT, requires identification of grammatical inaccuracies. Besides, it situates errors in a meaningful context and offers opportunities for multiple corrections of the same error. Overall, 311 participants took part in the study. A paired sample *t-test* showed a significant difference ($t = 24.10$, $p = .00$, $d = 1.91$) between the two tasks. Fifteen percent of the data (45 participants) was further inspected to have a better understanding of learners' error correction-pattern on the editing task. The results revealed that about 20% of the errors identified were wrongly corrected, which exposes limitations of the GJTs that only require judging the well-formedness of a construction. Moreover, performance on the editing task revealed a greater level of diversity in learners' responses. On average, eight of the total twelve grammatical features were corrected in 5.87 (*SD* = 2.47) different ways. The variety of responses on the ET reveals the dynamic nature of learners' interlanguage that allows for multiple ways for correction. The ET, as compared to the GJT, provides an ecologically more valid way to assess L2 learners' grammatical proficiency.

Keywords: grammaticality judgment tasks, reliability, scope

Introduction

One of the primary aims of the second language (L2) research is to describe L2 learners' linguistic knowledge. To accomplish this, various instruments are used, grammaticality judgment tasks (GJTs) being one of the most commonly used measurement tools. For example, a recently published meta-analysis reports 302 studies using a judgment task (Plonsky, Marsden, Crowther, Gass, & Spinner, 2019. Another meta-analysis examining only age effects on L2 grammar knowledge reports 20 studies that used some type of GJT for data collection (cf. Qureshi, 2016). Several explanations for the use of GJT have been put forth. For example, the GJTs are considered convenient in administration and analysis (Lowen, 2009; Plonsky et al., 2019), rigorous in determining the well-formedness of a construction (Riemer, 2009), and helpful in isolating and eliciting evidence about what is ungrammatical (Schütze, 1996). In L2 research, GJTs are used for various purposes, such as assessing theoretical claims, determining the type of L2 knowledge (i.e., implicit/explicit, procedural/declarative), but mainly for determining L2 learners' linguistic knowledge. Several researchers consider the use of GJTs as a reliable means for assessing learners' knowledge of structures and rules in a second language (cf. Bley-Vroman, Felix, & Ioup, 1988;

Chaudron, 1983; Gass 1994; Leow, 1996; Mandell, 1999). However, others raise concerns about their effectiveness in tapping into L2 learners' actual linguistics repertoire (Alanazi, 2015; Birdsong, 1989; Ellis, 1991; Sorace, 1985; Schütze, 2011; Tabatabaei & Dehghani, 2011). The questions raised about the use of GJTs include the effects of behavioral and extra-grammatical influences. For example, learners may "guess if they are not sure, balance between the number of sentences they judge grammatical and deviant, [and] they may avoid judging more than a certain number of consecutive sentences as ungrammatical" (Ellis, 1991, p. 164). Moreover, the extra-grammatical factors, such as, processing constraints, sentence complexity, and semantic irregularity (Birdsong, 1989; Ellis, 1991) may influence to mask what learners base their judgments on, and whether, they focus on the targeted error while making a judgment is unclear.

Moreover, the nature of L2 knowledge elicited through a GJT is also deliberated as limited in its scope. The scope of the GJT – as used in this research – concerns its ecological validity and subtractive nature. It has been argued that GJTs offer an unnatural paradigm for assessing linguistic knowledge (Plonsky et al. 2019) as the sentences provided are usually isolated – devoid of any meaningful contextual information (Qureshi, 2018). Moreover, their predominant dichotomous nature, which requires L2 learners to judge a sentence either as grammatically acceptable or unacceptable, appears subtractive – focusing mainly on what learners might be unable to do. Alternatively, a more additive paradigm should be permissive enough to allow learners to offer any correction, which would be contextually appropriate. The following section provides an overview of research on GJT concerning its reliability and scope.

Literature Review

Reliability

While validity can be generally defined as the extent to which a test measures what it claims to measure, reliability concerns how well does a test measure what it claims to measure (Cameron, 2001). For example, if a test *A* claims to measure *John's* knowledge of past tense, scores on the exam should precisely reflect John's familiarity with the tense. In case John retakes the same test, or if another grader scores the test, John's performance should generally appear consistent, without much change. The reliability of a data collection task is vital as the performance on the task is indicative of participants' knowledge in the language domain assessed by the task.

GJTs, mainly used for assessing L2 learners' grammatical knowledge, have been frequently debated for their reliability and generalizability. Concerns are often raised whether scores on a GJT are a reliable measure of L2 grammar proficiency and whether there exists any relationship between scores obtained on a GJT and other tasks that measure grammar knowledge. To determine the dependability of GJTs, researchers have predominantly applied two study-designs, which include (a) test-retest, involving two administrations of the same GJT at two intervals, and (b) comparison of participants' performance on a GJT viz-a-viz another task. In a few cases, learners' performance on a GJT in their second language is also compared with performance on a GJT in their L1 or against the performance of native speakers.

GJTs: Test-retest Design

The studies involving test-retest design generally reveal GJTs as a less reliable measure of assessing L2 proficiency in grammar and suggest caution in interpreting findings. Ellis (1991) conducted one of the earliest investigations of the reliability of GJTs using a test-retest design. Participants in his study (21 Chinese learners of English as an L2) were first administered a standard GJT, then after a week, a subgroup of the total sample (i.e., 8 participants) was asked to retake a shorter version of the actual GJT, now containing 10 out of the original 40 items. Finally, this subgroup of eight participants was requested to complete a think-aloud protocol. Results revealed a higher level of within-subject inconsistencies from one to the other administration of the judgment task. This led Ellis (1991) to conclude that "learners' judgments can be inconsistent, and, therefore, unreliable, when they are unsure" (p. 181).

A more recent study following the test-retest design was carried out by Alanazi (2015). The study involved 36 Saudi EFL learners who had completed a year-long foundational English course and had been in the UK for two years. Participants were administered a GJT involving past tense in three different planning conditions: guided, semi-guided, and no guidance, and were administered the GJT twice with an interval of eight weeks in between. The planning conditions did not reveal any significant difference at the first administration of the GJT, while at time two, a significant difference emerged in favor of the guided group. This change in the guided group's performance at time two led Alanazi (2015) to conclude that the GJT did not reflect the actual proficiency of the learners involved in the study.

Another study involving a similar test-retest design was carried out by Tabatabaei and Dehghani (2011). The study administered a verb-complementation task to 30 advanced EFL learners enrolled in a graduate program in Iran. The GJT was timed and computerized and was conducted twice with an interval of two weeks in between. The test-retest examination revealed a low correlation $r(30) = .399$, $p < 0.05$ (p. 178) between learners' scores on the same task. The inconsistent results of the study indicate GJT as a less dependable measure for mapping L2 learners' grammar knowledge.

However, unlike most other studies following a test-retest design, the results of Gass (1994) indicate GJTs as generally reliable measures of L2 learners' grammar knowledge. Gass (1994) also examined the test-retest reliability of GJTs that focused on relative clauses. Her study involved 21 Chines EFL learners who had completed a required English foundation program. The participants were administered a GJT twice, with a one-week interval. The learners were asked to make a binary judgment on 30 sentences first and then rate their degree of confidence about the judgments on a seven-point scale. The results of her study revealed a consistent performance in the two administrations of the GJT with reliability coefficients of .59 and .64 in binary and scaler judgments, respectively. However, about 19 % and 43 % of variation was also observed in learners' binary and scaler judgments. This led Gass (1994) to conclude that the reliability of GJTs is indivisible from learners' indecisiveness about the exact nature of sentences due to their partial knowledge of certain aspects being assessed, and GJTs "are indeed reflective of patterns of second-language use" (p. 320).

Nonetheless, generally speaking, most studies using a test-retest design reveal inconsistent findings in the two administrations of a GJT. This is also confirmed by studies that compared learners' performance on a GJT in L2 with their L1 or against the performance of native speakers. The findings of these studies also reveal varying outcomes. For example, in Davies and Kaplan (1998), adult L2 learners, and in Liceras (as cited in Mandell, 1999), native Spanish speakers showed variation in their performance on a GJT from one administration to another.

GJTs: Comparison Design

The set of studies that have adopted a comparison design have predominantly used a production task (i.e., oral or written) for comparison. These studies generally approve GJTs as reliable measures for assessing L2 grammar knowledge. For example, Leow (1996) compared L2 learners' performance on a GJT with two production tasks (i.e., oral & written). Participants in the study were 30 undergraduate students enrolled in the first-semester Spanish course. The learners were administered a GJT and two production tasks, focusing on the noun-adjective agreement in Spanish. The tasks were administered, first, in week three, and then, in week fourteen. The GJT task required the participants to identify a sentence as either correct or incorrect, locate the error, and explain a grammatical rule. For the oral and written production tasks, learners were shown a series of drawings and sets of questions and were requested to produce sentences answering the questions. The results revealed a strong and significant relationship between the GJT and the written production task and a moderate but significant relationship between the GJT and the oral production task at both administrations.

Mandell (1999) compared L2 learners' performance on a GJT and a Dehydrated Sentence Test (DST). While the GJT necessitated the participants to make the typical binary judgments, the DST required the participant to construct a sentence by using a set of words provided by the researcher. The participants involved 91 learners enrolled in the second, fourth, and fifth semesters of Spanish L2 course. The main grammatical feature investigated involved v-movement in three constructions: (a) wh-questions, (b) yes/no questions, and (c) adverb placement. Similar to Leow (1996), the results in Mandell's study displayed a strong and significant relationship between learners' scores on the GJT and the DST across all three levels.

However, Christie and Lantolf (1991) do not report a similar pattern of a strong correlation between a GJT and other production tasks. They investigated English L1 speakers' knowledge of the Pro-drop parameters in Italian, comparing learners' performance on a GJT with an oral narrative task. Observing variation in a learner's performance on the two tasks, they concluded that "judgment data may not necessarily inform us of a learner's changing interlanguage grammar" (p. 18). Nevertheless, their findings have limitations as the second task (i.e., oral narrative task) was performed by only two learners, and the inconsistent finding reported is based on the performance of one of the two subjects who completed the second task.

To summarize, the studies reported in the two designs discussed above generally present mixed findings. Most previous research adapting test-retest design report inconsistencies in learners' performance when the same judgment task is conducted on two or more occasions. In contrast, the studies that assess L2 learners' performance on a GJT against another task, typically indicate GJTs as reliable measures of examining L2 knowledge.

GJTs: Scope

Like the reliability, researchers have also argued about the nature of L2 knowledge as static or dynamic (e.g., Larsen-Freeman, 2005). The dynamic perspective stresses the role of the environment and interaction in second language processing. It castoffs the reflexive access to errors; instead, it emphasizes the role played by individuals in the context of meaningful interaction (Shanker & King, 2002). Emphasizing the role of context, Larsen-Freeman (2002) points out that "language, or grammar, is not about having; it is about doing: participating in social experiences" (p. 42). Despite the importance of authentic context in language learning and assessment, most GJTs are administered stand alone. Plonsky et al., (2019), reports that, although in some cases GJTs accompany some contextual information, such as embedding a GJT item with a story or a picture, in a majority of cases, (i.e., 81%), no contextual information is provided. Unlike the typical decontextualized GJTs that require learners to make grammaticality judgments, a task based on the dynamic approach to second language learning would allow learners a greater opportunity to engage with the material in an approximate real context, which is currently lacking.

Another issue about the scope of knowledge assessed through GJTs concerns the way learners are asked to respond to items in the GJT task. For example, participants may be asked to judge between grammatically correct and deviant sentences only or to locate and correct errors, or in some cases, they may be asked to explain the errors (Ellis, 1991). Although the practice of asking learners to identify or correct errors exists in previous research, it is not very common. Plonsky et al. (2019) point out that only 23% of the research papers in their meta-analysis of 302 studies using judgment tasks employed these methods, while the majority depended only on the intuitive judgments. This over-reliance on the intuitive judgments appears problematic as it is unclear what learners base their judgments on, or even if they identify the error contained in the sentences.

Moreover, the majority of studies that depend on intuitive judgments assume the possibility of only one correct form because no additional alternatives are provided. This approach is theoretically and empirically wrong because the same error might be corrected in several ways. To illustrate, two learners might differently correct the same error, or the same learner may exhibit variation in the correction of the same feature on various occasions. Most GJTs that involve intuitive judgments ignore the variation inherent in L2 learners' actual language use. Hence, to examine the dependability of a GJT, there is a need to compare the results obtained from it with another task that situates errors in an authentic context and offers opportunity for variation in corrections. The current study attempted this; a brief description of the study follows.

The Current Study

To investigate the dependability of a GJT against a more authentic measure, the current study used an editing task. The editing task was chosen for several reasons. First, the editing task allowed the insertion of violations for the same twelve grammar rules that were contained in the GJT used in the current study. This may not have been equally possible with other measures (e.g., written or oral productions); hence, a limitation of such studies. Second, contrary to the binary judgment tasks that only require L2 learners to make an intuitive judgment about the grammaticality of a construction, the editing task used in this study provided an opportunity for error identification as well as correction. Hence, it provided more reliable data about learners' grammar knowledge. Some might argue that the same procedure could be utilized with a GJT. Several studies (e.g., Ellis, 1991; Tabatabaei & Dehghani, 2012) have done this in the past. Although this is true, identifying and correcting errors in a GJT may not provide as authentic a context as an editing task, which situates grammatical violations in a meaningful setting. L2 learners in real life are often confronted with editing their own and their peers' papers that do involve error identification and correction; it may be hard to argue that the GJTs also provide a similar authentic context. Last, an editing task provided an opportunity to correct the same error in multiple ways, which reflects an ecologically more authentic approach. Hence, by offering a meaningful context and the opportunity for identifying and correcting errors in multiple ways, the editing task used in the study appears to better align with the dynamic perspective of language development.

Methods

Research questions

1. To what extent does L2 learners' grammatical knowledge, as reflected on a GJT task, matches with their performance on an editing task (ET)?
2. To what extent do L2 learners' response types on the editing task (e.g., identified but corrected wrongly, ICW; and identified and corrected, IC) correspond to each other and with their performance on the GJT task?
3. To what extent does the editing task reveal diversity in L2 learners' corrections for the same grammatical errors?

Participants

Three hundred and eleven L2 learners participated in the study. The current study involved three questions. For questions 2 and 3, data from only 45 participants were further analyzed. Details about the participants' background are provided in Table 1.

TABLE 1
Participants Background Information

Overall Analysis	
N	311
Minimum exposure to English	5 years
Age at Testing	
M	22.55 (2.7)
Range	17 - 47
Educational level	
Undergraduate	105 (32 %)
Master's	224 (68 %)
Not provided	6
Self-rated proficiency in English	3 (out of 5)
Sub-analysis for Q2 & 3	
N	*45*
Minimum exposure to English	5 years
Age at Testing	
M	22.85 (2.2)
Range	18 - 30
Educational level	
Undergraduate	11 (24 %)
Master's	34 (74 %)
Self-rated proficiency in English	3 (out of 5)

The participants, on average, had at least five years of exposure to English as a foreign language (FL) in an instructed context. Their mean age at the time of data collection was approximately 23 years. A majority of learners were enrolled in a master's degree program.

Background Questionnaire

A background questionnaire consisting of 16 items was administered for collecting information about participants' age at testing, years of exposure to English in the instructed context, their educational background, and self-rated proficiency in English.

Grammaticality Judgment Task

The grammaticality judgment task by DeKeyser (2000) with adaptations made by Qureshi (2018) was used in the study. This instrument was initially used in a highly influential work of Johnson and Newport (1989). Since then several studies have adapted this for assessing L2 learners' grammar knowledge (e.g., Seol, 2005; DeKeyser, Alfi-Shabatay, & Ravid, 2010). The GJT involved violations of twelve grammatical features considered as problematic for L2 learners. These grammatical features contained in the GJT were *past tense, plural, third-person singular, present progressive, determiners, pronominalization, particle movement, infinitives, gerunds, yes/no- and wh-questions.* For each rule violation, two pairs of grammatical and ungrammatical sentences were included in the instrument. In some cases, for example, word order, third-person singular, and plurals, more than one rule were violated; hence, two pairs of sentences per each violation type were included in the instrument. All the items in the GJT were randomized. To ensure the readability level of the sentences does not pose difficulty in sentence processing and subsequently impact grammatical judgment, lexile measure of the sentences was examined using a freely available application *Lexile Analyzer* (2020). The sentences in the GJT obtained a lexile range of (10L to 270L), indicating the difficulty level of the sentences well below the third grade level.

The average length of the sentences was 7.5 (*SD* = 1.6). The resultant GJT contained 114 items and obtained the Kuder- Richardson 20 (KR-20) reliability coefficient of .89. The GJT conducted with the sub-sample of 45 participants obtained a reliability coefficient (KR-20) of .83. The GJT was paper-based and non-pressured – the participants were given five extra minutes than the time spent during the piloting phase.

Editing Task

To validate the GJT, an editing task was used. The task was adapted from Qureshi (2018). The passage described the effects of Wars on health and poverty in Africa in very simple and accessible language, without using lexis and concepts that might cause accessibility issues. The ET contained 229 words and 24 morphosyntactic errors violating the same 12 grammatical rules as contained in the GJT. Two errors for each grammatical rule were inserted in the ET. The author (XXX) reported the task readability as 880, which represents a text used at the fourth-grade level*,* which appears very close to the readability level of the GJT at the third grade level. Moreover, similar to the GJT, the editing task was also loosely-timed and administered on paper. The average length of the sentences was 13.12 (*SD* = 4.1). Although sentences in the ET appear slightly longer than those contained in the GJT, this should not have caused additional strain as the sentence difficulty was at a very accessible level (i.e., 4th grade), and the task was not speeded. A description of errors and their potential corrections are provided in Table 2. Table 2 contains only a few possible correct forms for each violation, while in the actual performance, learners came up with several variants of the corrections that seemed situationally correct in the context. Similar to the GJT, the ET was conducted as paper-based, non-pressured, and it obtained the Kuder- Richardson 20 (KR-20) reliability coefficient of .83. The ET conducted with the sub-sample of 45 participants obtained a reliability coefficient (KR-20) of .83 as well.

TABLE 2
Grammatical Features, Error Types, and Potential Corrections

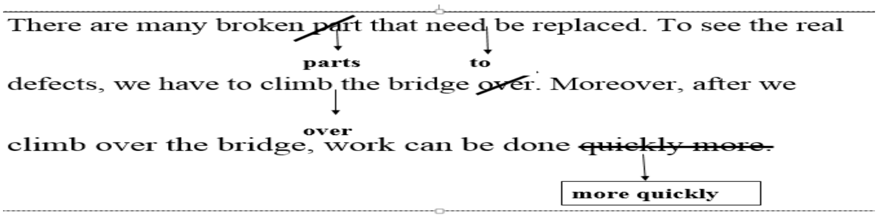
Features	**Error and Correction**		**Error and Correction**	
Past tense	Begin	(began)	have taken	(took)
Plural	part	(parts)	tradition	(traditions)
Third-person	tooks	(took)	cause	(causes)
Present Progressive	become	(ing/have become)	grow	(growing/have grown)
Determiners	traditional	(the traditional)	virus	(the virus)
Pronominalization	future	(their future)	her	(their)
Particle Movement	fighting control over	(fighting over control)	looking . . out	(looking out/looking)
Gerunds	face	(facing)	continue	(ing/ed)
Infinitive	take	(to take)	to solving	(to solve)
Aux. wrong/omitted	were not	(did not)	not	(do not)
Word Order	facing today Africa	(facing Africa today)	rapidly has	(has rapidly)
Adverb	dramatic	(dramatically)	quick	(quickly)

Procedures

The participants were administered the tasks in the following order: (1) GJT, (2) ET, and (3) the background questionnaire. As completing the ET first could sensitize learners about the errors in the sentences, it was conducted after the GJT. The BQ was administered at the end, assuming that it would require relatively easily accessible information.

The GJT required binary judgments of grammatically correct and deviant sentences. While administering the GJT, the participants were first read directions and then shown two sample practice items on the screen. After they felt comfortable with the procedure, they were allowed to complete the task. On average, it took participants about 35 minutes to complete the GJT.

Similar to the procedures for the GJT, the participants were first told directions for the ET and then shown a sample text depicting the intended correction procedure. The following example was shown to the participants before they started the task.



There are many broken ~~part~~ that need be replaced. To see the real
parts to
defects, we have to climb the bridge ~~over~~. Moreover, after we
over
climb over the bridge, work can be done ~~quickly more.~~
more quickly

The learners were instructed to perform the following on the ET: (a) cross an error and replace it with the correct form, (b) rearrange word order where needed (which could also involve crossing out a word/s), and (c) insert a missing word.

Analysis

For investigating research question 1 that attempted to validate L2 learners' grammatical knowledge as reflected on the GJT with their performance on the editing task, overall scores on the ET were compared with learners' performance on the GJT. The responses on the GJT were coded as 0 and 1 for the incorrect and correct judgments, respectively. Similarly, for the ET, all the responses that were identified, irrespective of whether these were corrected or not, were coded as 1, and those not identified, as 0.

For question 2, a sub-sample of 45 participants from the total data was randomly chosen. The GJT was coded dichotomously: incorrect, 0, and correct, 1, but for the ET, now data were coded only for those responses that were correctly identified. These responses were further coded for three conditions: (a) identified but not corrected (b) identified but corrected wrongly (ICW), and (c) identified and corrected (IC). Each category, if confirmed, was coded as 1. While analyzing data, only four instances of "identified but not corrected" were observed; hence, this category was merged with 'identified but corrected wrongly'. These categories were combined because these reflected a more similar concept (i.e., identified only) as compared to the third construct that examined 'identified and corrected'. The resultant data for the ET represented two conditions after coding: (a) ICW, involving error identification, but wrong correction and (b) IC, comprising errors that were identified as well as corrected.

As the questions 1 and 2 examined the same participants' performance on two different tasks, a paired sample *t*-test was run. Before conducting the t-test, scores on the two tasks were normed at 100. The assumptions regarding the continuousness of data and normalcy of distribution were checked and met. For question 3 that explored the variety of correct responses on the editing task, simple frequency counts were computed.

Results

To examine whether L2 learners' grammatical knowledge, as reflected on the GJT task, would be validated by their performance on the editing task, a paired sample *t-test* was computed. The results of the analysis are provided in Table 3.

TABLE 3
Paired Sample t-test for Grammaticality Judgment Task and the Editing Task (Based on Percentage Correct)

Variables	*n*	*Mean*	*SD*	*t*	*df*	*p*	*d*
GJT	311	69.83	11.18	*24.10	310	.01	1.91
Editing Task	311	38.65	20.11				

*Equal variance not assumed

Results of the paired sample *t-test* revealed a significant difference between scores on the grammaticality and editing tasks ($t_{310} = 24.10$, p =.01). The results indicated that learners, on average, obtained approximately 31 points higher on the GJT as compared to the ET (95% CI [28.32, 33.36]). Further analysis also revealed a moderate and significant correlation (r = .04, p = .05) between learners' scores on the grammaticality and the editing tasks for the whole group analysis.

The second research question inquired whether the ET would reveal differences in the levels of grammatical knowledge (i.e., identified but corrected wrongly and identified and corrected), and whether these conditions would correspond with learners' performance on the GJT. To answer this question, three iterations of a paired sample *t*-test were run. Afterward, learners' performance on the ICW and IC were also assessed through percentage and ratio analysis. To answer the second research question, three comparisons were made in the following order: a) ICW and IC, b) ICW and GJT, and c) IC and GJT. Details of the comparison are provided in Tables 4 and 5.

TABLE 4
Paired Sample t-test for Grammaticality Judgment Task and the Two Conditions of the Editing Task (IO and IC)

Response type	***n***	***Mean***	***SD***	***t***	***df***	***p***	***d***	***Lower***	***Upper***
								95%	
ICW	45	8.00	6.12	-9.336	44	<.01	-2.00	-11.01	-06.53
IC	45	41.84	23.04						
ICW	45	8.00	6.12	12.31	44	<.01	-7.00	25.70	35.76
GJT	45	81.35	12.86						
IC	45	41.84	23.04	16.57	44	<.01	2.11	34.70	44.31
GJT	45	81.35	12.86						

TABLE 5
Correlations between the Scores on the GJT and the Two Conditions of the ET

Tests	IC	GJT
ICW	-.08* (.60)	-.03 (.84)
IC	--	.74 (.00)
GJT	--	--

Results of the paired sample *t-tests* revealed a significant difference between the two conditions of the ET: ICW and IC ($t_{44} = -9.33$, $p < .01$), and each of these conditions and the GJT: ICW and GJT ($t_{44} = 12.31$, $p < .01$), and IC and GJT ($t_{44} = 16.57$, $p < .01$). For the correlation analysis, scores on the two conditions of the ET (ICW and IC, $r = -.08$, $p = .60$), and between the ICW and GJT ($r = -.03$, $p = .84$) were not significantly correlated. This finding indicates that the ICW could be a potential source of measurement error in the GJT, which could be eliminated by using an editing task. However, there existed a strong, significant, and positive correlation between participants' scores on the IC and GJT ($r = .74$, $p = .001$). This significant correlation between the 'identified & corrected' condition of the editing task and the GJT indicates that the grammaticality judgments on the GJT may not involve merely 'intuitive judgments', rather learners might immerse into error identification much like they do in an editing task.

To further explore the cases where learners could successfully identify an error but corrected it wrongly, raw counts and percentages for 'identified and corrected' and 'identified but corrected wrongly' conditions were computed. The results of the analysis are provided in table 6.

TABLE 6
Raw Counts and Ratios for The Two Conditions of the Editing Task

n	*IC*	*ICW*	*Ratios ICW: IC*	*Percentage ICW & IC*
45	453	96	1:5	20 to 80

Overall, there were 1,080 instances where learners could identify and correct an error (24 errors in the task * number of participants 45 = 1080). The analysis unveiled that, overall, there were 453 occurrences where learners identified and corrected errors, and 96 instances where learners identified the errors but offered wrong correction. The wrong corrections made about 18% of the total errors identified, which made approximately 1:5 ratio of the errors identified.

Questions 3 explored the diversity in L2 learners' corrections for errors in the editing task. To answer this question, simple frequency counts were computed. Some grammatical features, for example, past tense, plurals, third-person singular, and determiners, showed a consistent pattern without any variety in corrections. However, other features displayed a greater variation in the provided corrections. Table 7 presents the number of ways errors pertaining to various grammatical features were uniquely corrected.

TABLE 7
Number of Unique Ways Errors Were Corrected on the Editing Task

Features	**Number of different ways Corrected**
Present Progressive	4
Pronominalization	4
Particle Movement	5
Gerund or Present participles	8
Infinitive	4
Aux. wrong/omitted	6
Word Order	11
Adverb	4
Mean	5.87 (*SD* = 2.47)

A few examples, demonstrating how ingenious learners can be when required to correct errors, are presented in table 8. These support a dynamic approach to language learning, which is in sharp contrast to the static perspective expecting relatively fixed outcomes as obtained through the binary judgments.

TABLE 8
Examples of Variations in Corrections for the Same Grammatical Error

Present Progressive

Sentence with **errors**

They are unable to take any measures for a better future, as a result many countries **are become poorer** and their **problems are grow**.

Corrections

They are unable to take any measures for a better future, as a result many countries ***are becoming poorer*** and their problems ***are growing.***

They are unable to take any measures for a better future, as a result many countries ***are poor/er*** and their problems are growing**.**

They are unable to take any measures for a better future, as a result many countries ***have become poorer*** and their problems ***have grown***.

Particle Movement

Sentence/s with errors

Many of the problems facing Africa today have been worsened by **fighting** control **over** of the government

Young and talented Africans are **looking** to the rest of the world **out**.

Corrections

Many of the problems facing Africa today have been worsened by ***fighting over control*** of the government

Many of the problems facing Africa today have been worsened by **fighting for control** of the government

Young and talented Africans are ***looking out to*** the rest of the world.

Young and talented Africans are ***looking to*** the rest of the world ~~out.~~

Young and talented Africans are ***looking at*** the rest of the world.

Gerund or Present participles

Sentence with errors

The biggest **problem face Africans today** is **the continue threat** of wars.

Corrections

The biggest problem ***facing*** Africans today is the continuing threat of wars.

The biggest problem ***faced by*** Africans today is the continuing threat of wars.

The biggest problem Africans ***face*** today is the continuing threat of wars.

The biggest problem ***which Africans face*** today is the continuing threat of war.

The biggest problem, ***which Africans are facing*** today, is the continuing threat of war.

The biggest problem facing Africans today is the ***continuing*** threat of wars.

The biggest problem facing Africans today is the ***continued*** threat of wars.

The biggest problem facing Africans today is the ***continuous*** threat of wars.

Auxiliaries

Sentence with errors

They **not understand** much about African tribal traditions, that's
why the borders of these **countries not** match the traditional borders.

Corrections

They ***did not understand*** much about African tribal traditions.

They ***were not informed*** about African tribal traditions.

They ***were not aware of*** African tribal traditions.

They ***are not able to*** understand African tribal traditions.

They did not understand much about African tribal traditions, that's why the borders of these ***countries do not match the*** traditional borders

They did not understand much about African tribal traditions, that's why the borders of these ***countries did not match*** the traditional borders.

Word Order

Sentence/s with errors

Many of the **problems facing today Africa** have been worsened by fighting over control of the government.
HIV, the virus that causes AIDS, **rapidly has** spread in Africa.

Corrections

Many of the problems ***facing Africa*** have been worsened by fighting over control of the government.
Many of the problems ***facing Africa today have*** been worsened by fighting over control of the government.
Many of the problems ***Africa faces today*** have been worsened by fighting over control of the government.
Many of the problems, ***which Africa is facing today,*** have been worsened by control of the government.
Many of the problems ***Africa is facing today*** have been worsened by fighting over control of the government.
Many of the problems ***facing Africa nowadays*** have been worsened by fighting over control of the government.
HIV, the virus that causes AIDS, ***has rapidly*** spread in Africa.
HIV, the virus that causes AIDS, ***has spread rapidly*** in Africa.
HIV, the virus that causes AIDS, ***has spread*** in Africa ***rapidly***.
HIV, the virus that causes AIDS, ***is rapidly spreading in*** Africa.

Discussion

The results of the study indicate three main conclusions. First and probably the most valuable contribution of the current study is reflected in the provision of the ET as an authentic and meaningful task. Previous research has emphasized the role of the environment in meaningful interaction (cf. Shanker & King, 2002). Unlike the typical GJTs that present binary options in decontextualized contexts, the editing task allowed the participants to rectify errors in any way they deemed contextually appropriate, which is a more realistic expectation in authentic contexts and aligns well with the dynamic perspective of language learning. The dynamic perspective of language learning situates a learner in a realistic social context and views the process of language learning as participation in social experiences (Larsen-Freeman, 2002, p. 42). The ET in the current study provided learners an opportunity to interact with a meaningful text that discussed poverty and war in Africa. It also allowed learners to identify and correct errors in a context that they are more likely to experience in reality – editing their own and their peers' writing - than a typical GJT that elicit responses to decontextualized errors.

Second, the use of ET appears to be a viable solution to the limitations expressed for the GJT. Previous research indicates several limitations of the typical GJTs. For example, some researchers express concerns that while making judgments, learners may not focus on the actual errors contained in the sentences; instead, they may be influenced by behavioral (i.e., they may guess, keep balance; Ellis, 1991; be biased and over-reject; Birdsong, 1989) or external factors (e.g., complex structure or semantic difficulty in a sentence; Ellis, 1991). The findings of the current study support these concerns by indicating a low and non-significant correlation (r = -.03, p = .84) between the GJT and the 'identified but corrected wrongly' category on the editing task. It can be argued that the ICW could be viewed as the source of measurement error, which was mixed in the GJT but could be excluded by using the ET.

Third, concerns the predictive ability of GJTs for general language proficiency. Previous research suggests examining the predictive ability of the typical GJTs for second language learners' general language proficiency as assessed through various language proficiency tests, such as TOEFL and IELTS (Ellis, 2005). Although the current study did not compare performance on the GJT against a standardized language proficiency test, the editing task used in the study reflected general language use, at least, as authentically as other language proficiency measures. Hence, the findings of the study bear relevance here. The significant and positive correlation between the GJT and editing tasks used in the current study posits that both the instruments measured a similar construct – grammatical ability. However, a significant average difference between scores on the two tasks indicates that the participants' performance on the GJT is not indicative of their ability to correct errors on an editing task, which they may be required in their actual instructed language learning settings. The findings of the current study indicate that learners might be unable to rectify 1 out of every 5 errors they can identify. Hence, findings based on GJTs that only require learners to make judgments about grammaticality without error correction may report average scores, which are inflated by 20%. Thus, it might be concluded that the typical GJTs have limited predictive ability for general language proficiency in instructed contexts.

Future Directions

The current study investigated L2 learners' knowledge of twelve grammatical rules. Future studies might consider reducing the number of grammatical features to only those that have been reported as problematic for the population sample under investigation. Moreover, items per each rule type were set up to only two pairs of grammatical and ungrammatical sentences. In case of the editing task, the number of items was reduced to only two per each grammatical feature. In the future, studies should include a greater number of items for each grammatical feature to better account for the reliability of the instrument (DeKeyser, 2013). Moreover, the instruments used in this research were loosely-timed and visual in nature, so any implication for scenarios employing timed and aural measures should be drawn with caution.

Conclusions

This paper, comparing performance on a GJT with an ET, points out a need for contextually more authentic tasks. Authentic measures, like the editing task reported in this paper, offer a greater opportunity for exploring learners' language proficiency in meaningful contexts on the one hand, while on the other, these may offer opportunities for a variety of possible corrections; otherwise, impossible in a binary-judgment scenario. Moreover, participants' 'identified but wrongly corrected' responses on the editing task indicate a potential source of error in the GJTs that only require binary judgments of grammaticality without requiring error identification or correction. Plonsky et al. (2019) point out that about 77% of judgment used in L2 research adopted this approach. Lastly, the variety of corrections on the editing task supports an additive approach to second language learning, where the focus is not only about what learners cannot do with language, but what they can, and in how many different ways.

References

Alanazi, S. (2005). The validity of grammaticality judgment task on Saudi EFL learners. *International Journal of Applied Linguistics & English Literature, 4*(6), 78–83.

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

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

Bley-Vroman, R. W., Felix, S. W., & loup, G. L. (1988). The accessibility of Universal Grammar in adult language learning. *Interlanguage Studies Bulletin (Utrecht)*, *4*(1), 1–32. https://doi.org/10.1177/026765838800400101.

Cameron, L. (2001). *Teaching language to young learners.* Cambridge: Cambridge University Press.

Chaudron, C. (1983). Research on metalinguistic judgments: A review of theory, methods, and results. *Language Learning, 33*, 343–377.

Christie, K., & Lantolf, J. (1991). The ontological status of learner grammaticality judgments in UG approaches to L2 acquisition. *Paper presented at Second Language Research Forum,* Los Angeles.

Davies, W.D., & Kaplan, T.I. (1998). Native speaker vs. L2 learner grammaticality judgments. *Applied Linguistics, 19* (2), 183-203.

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*, 499–533.

DeKeyser, R. (2013). Age effects in second language learning: Steppingstones toward better understanding. *Language Learning, 63,* 52–67.

Ellis, R. (1991) Grammaticality judgments and second language acquisition. *Studies in Second Language Acquisition, 13*, 161–186.

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

Gass, S. (1994). The reliability of second-language grammaticality judgments. In E. Tarone, S. Gass, & A. Cohen (Eds.), *Research methodology in second language acquisition* (pp. 303-322). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hayes, B. P. (2000). Gradient well-formedness in Optimality Theory. In J. Dekkers, F. van der Leeuw, & J. van de Weijer (Eds.), *Optimality theory: Phonology, syntax, and acquisition* (pp. 88–120). Oxford University Press, Oxford.

Larsen-Freeman, D. (2005). Second language acquisition and the issue of fossilization: There is no end, and there is no state. In Z.-H. Han & T. Odlin (Eds.), *Studies of fossilization in second language acquisition* (pp. 189–200). Clevedon, UK: Multilingual Matters.

Leow, R. P. (1996). Grammaticality judgment tasks and second-language development. *Georgetown University Round Table on Languages and Linguistics,* 126–139.

Lexile Analyzer (2020). The lexile framework for reading. Retrieved from https://www.lexile.com/ analyzer.

Loewen. S. (2009) Grammaticality judgment tests and the measurement of implicit and explicit L2 knowledge. In R. Ellis, S. Loewen, & C. Elder (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 94–112). Bristol: Multilingual Matters.

Mandell, p, B. (1999). On the reliability of grammaticality judgement tests in second language acquisition research. *Second Language Research, 15*(1), 73–99.

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2019). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research.* https://doi.org/10.1177/0267658319828413

Qureshi, M. A. (2016). Age and knowledge of morphosyntax in English as an additional Language: grammatical judgment and error correction. Retrieved from *The International Research Foundation (TIRF-report)* website: http://www.tirfonline.org/wpcontent/uploads/2016/08/TIRF_DDG_2014_MuhammadAsifQureshi_Final.pdf

Qureshi, M. A. (2018). Age and knowledge of morphosyntax in English as an additional language: Grammatical Judgment and Error Correction. *International Review of Applied Linguistics.* DOI https://doi.org/10.1515/iral-2015-0062

Riemer, N. (2009). Grammaticality as evidence and as prediction in a Galilean linguistics. *Language Sciences, 31*, 612–633.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistics methodology*. Chicago: The University of Chicago.

Shanker, S., & King, B. (2002). The emergence of a new paradigm in ape language research. *Behavioral and Brain Sciences, 25*(5), 605–656.

Sorace, A. (1985). Metalinguistic knowledge and language use in acquisition poor environments. *Applied Linguistics, 6*(3), 239–254.

Tabatabaei, O., & Dehghani, M. (2011). Assessing the reliability of grammaticality judgment tests. *Procedia - Social and Behavioral Sciences, 31*, 173–182.