Assessing the reliability of grammaticality judgment tests

Omid Tabatabaei a *, Marzieh Dehghani a

a English Department, Najafabad Branch, Islamic Azad University, Najafabad, Iran

Abstract

Second language acquisition researchers have been using Grammaticality Judgment Tests (GJTs) since the mid-1970s in order to assess the linguistic competence of second language learners in their L2. A number of researchers (e.g., Gass, 1994; Ellis, 1991) have raised serious questions concerning the reliability of this type of test as a measure of L2 learners’ linguistic competence. The purpose of this study was to examine the reliability of GJTs used in foreign language context and also to explore the relationship between timed GJTs and delayed GJTs. After administering a standard language proficiency test (OPT), 30 advanced out of a pile of 80 EFL learners were selected for this study. Participants were asked to make judgments about 34 sentences included in a computerized GJT. The grammatical structure chosen for this study was verb complements. After second administration of the same computerized GJT, various methods were used in order to examine the reliability of timed GJTs. The results of test-retest analysis and internal consistency reliability revealed that the GJT used in this study had low level of reliability. Moreover, the analysis of response patterns showed that participants were not stable in their judgments and also they were reluctant to use not sure response when they were uncertain. Therefore, their judgments did not exactly reflect their grammatical knowledge. And finally the relationship between timed GJT and delayed GJT was weak which indicated that participants may have used different types of knowledge under different tests administration conditions. The results of this study suggest that the GJT used in this study is not a reliable measure of EFL learners’ knowledge about verb complements and researchers should use this kind of test with more caution.

Keywords: Universal Grammar (UG); Grammaticality Judgment; Grammaticality Judgment Tests (GJT); Reliability.

1. Introduction

One of the most widespread data collection methods that linguists use to test their theoretical claims is GJTs (Tremblay, 2005). In these tests, learners are asked to make judgments about individual sentences; they should decide whether individual sentences are grammatical or ungrammatical. According to Rimmer (2006) “A standard method of determining whether a construction is well-formed is a grammaticality judgment test, where subjects make an intuitive pronouncement on the accuracy of form and structure in individual decontextualized sentences”, (P.246). GJTs have come to be used in second language acquisition research from the mid 1970s. It was due to the application of the Chomsky’s grammatical theories (e.g., Extended Standard theory and the Principles and Parameters theory) in many studies and also the emergence of this assumption that GJTs allow researchers to collect specific data about specific grammatical features in order to test their hypothesis, and also the assumption that data which are collected through GJTs are more representative of a learner’s competence in a particular language than natural occurring data (Davies & Kaplan, 1998). As Schütze (1996) states the use of GJTs in linguistic theory is necessary because firstly, these kinds of data collection methods provide the researchers with samples of

* Omid Tabatabaei.
E-mail address: tabatabaeiomid@yahoo.com

1877-0428 © 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Prof. Hüseyin Uzunboylu.
Open access under CC BY-NC-ND license, doi:10.1016/j.sbspro.2011.12.037
participants’ reactions to types of sentences that rarely occur in natural speech. Secondly, they gather negative evidence (evidence of what is ungrammatical) that natural language data does not contain such evidence. Thirdly, they distinguish production problems (e.g., slips, unfinished utterances, etc.) from grammatical production, and finally they minimize the influence of the communicative and representational functions of the language, therefore, they isolate the structural properties of the language.

According to many L2 researchers such as Bley-Vroman, Felix, and Ioup (1988), Chaudron (1983), and Gass (1994), L2 grammaticality judgment data are useful tools for researchers in order to investigate L2 learners’ competence (the abstract knowledge a speaker has of her language) separate from their performance (the actual use of language in concrete situation). In fact, in research design grammaticality judgment data are used to make judgment about learners’ linguistic competence which is the knowledge of syntactic structures and rules. They believe grammaticality judgment data reflect what the learners know not what they do.

In recent years, the use of GJTs in second language acquisition research design has become quite controversial (Riemer, 2009). One reason is due to the absence of clear criteria to determine the exact nature of grammaticality. There are questions related to grammaticality, whether it is a dichotomous concept or a gradient concept (e.g., Sorace & Keller, 2005). Another reason is questions related to the validity of grammaticality judgments, that is, the extent to which they actually reflect learners’ grammatical competence. In the past it was assumed that grammaticality judgment data are an exact reflection of L2 learners’ linguistic knowledge, and researchers such as Bley-Vroman et al. (1988), Gass (1994) and White (1989), argued that L2 grammaticality judgments provide valid data for L2 research; however, other researchers (e.g., Schachter & Yip, 1990; Birdsong, 1989; Ellis, 1991; White, 2003) suggested that grammaticality judgment data not only reflect linguistic knowledge of learners but also reflect other factors such as their processing constraints, response biases, and the nature of the target structures. Therefore, they do not provide a direct window into linguistic competence as it was assumed (Tremblay, 2005).

GJTs also have been criticized for their reliability, whether the data obtained from these kinds of tests are reliable or not. Some researchers such as Gass (1994) and Johnson, Shenkman, Newport, and Medin (1996) have examined questions related to the GJTs’ reliability. According to Birdsong (1989) L2 learners are often inconsistent in their performance on GJTs, and it is not clear what learners actually do when they make judgments and how the reliability of GJTs can be improved. Therefore, there is a need for researchers to recognize and investigate these problems and to find solutions.

Since there are still questions about the reliability of GJTs, the present study has intended to address the issue of the reliability of such tests.

2. Review of Literature

Reliability of GJTs in second language acquisition research has been a matter of concern for many researchers. Ellis (1991) is one of the first who employed a test-retest research design in his study to address the reliability of grammaticality judgments in second language acquisition. Ellis’s study had two phases, with one week interval between the two phases. In both phases of the experiment advanced ESL Chinese students were asked to make judgments about sentences involving dative alternation in English. In the second phase some of the participants were also asked to perform a think-aloud task. Based on the considerable inconsistency observed in his participants’ grammaticality judgments, Ellis suggested that “learners’ judgments can be inconsistent, and therefore unreliable, when they are unsure” (Ellis, 1991, p.181). He maintained that beginners are not suitable subjects for examining the reliability of GJTs because their judgment data are not validated by data from other types of tasks (e.g., oral production).

In the same vein, Gass (1994) examined the test-retest reliability of GJTs in order to investigate the reliability of such tests. In her study, College ESL learners took the same test twice with a one-week interval. Learners were asked to make binary judgments and then to rate the degree of their confidence about their judgments on a seven-point scale. The grammatical feature for this study was the function of relative pronouns. The test-retest reliability coefficients for overall sentences were .598 in the case of binary judgments and .644 in the case of
judgments on a rating scale. Learners changed 19.4 percent of their binary judgments and 42.6 percent of their judgments on a rating scale. Gass concluded that, based on her results, the issue of reliability of grammaticality judgments is inseparable from issue of indeterminacy, that is, learners’ incomplete knowledge or absence of knowledge of certain aspects of the L2 grammar.

In another study, Mandell (1999) compared data from GJTs with data from Dehydrated Sentence Tests (DSTs) (a slash-sentence test which is commonly used in the L2 classroom to examine L2 learners’ knowledge about word order) in order to investigate the reliability of GJTs. Data were collected from three levels (second, fourth and sixth semester) of adult L2 learners of Spanish. The results from the comparison of the two tests indicated that “a definite relationship existed between the standard GJT and the DST” And “the grammaticality judgments of L2 learners, although indeterminate, were consistent.” (Mandell, 1999, p.93). Therefore, Mandell concluded that GJTs were reliable measures of L2 learners’ linguistic competence.

Sometimes a sentence may pose processing problem for learners, in this case learners may judge a sentence as ungrammatical “due to properties of the comprehension process that are independent of grammatical knowledge” (Schütze, 2011, p. 211). According to Schütze (2011) one of the potential confound of GJTs to study linguistic competence is parseability. For example, linguists disagree about the grammatical status of sentences such as, if that John likes Mary surprises you, you obviously haven’t been paying much attention, while some claim that this sentence is grammatical some others judge it as ungrammatical. Thus, he suggests that grammaticality judgments may reflect processing factors rather than grammatical knowledge.

According to Birdsong (1989) there is also a need for researchers to examine response patterns of grammaticality judgments in order to investigate the reliability of GJTs. For example, Ellis (1991) observed that the judgments of the advanced and intermediate learners changed between 22.5 percent and 45 percent respectively. In addition to the variation across learners, the stability of responses changed depending on the structures tested and the response type (e.g., binary or preference). Sorace (1985) also found a nonsignificant and negative correlation between the judgment scores and the scores on the oral production task for beginners but a significant positive correlation for intermediate learners.

In a study by Bley-Vroman et. al. (1988) their participants judged sentences as ungrammatical more than as grammatical regardless of their grammaticality. They noticed that when the participants were uncertain, they tended to judge sentences as ungrammatical. However, it is not yet clear what causes such response biases. In their study, participants also received higher scores in ungrammatical sentences. But in other studies (e.g., Ellis, 1991; Hawkins, Towell, & Bazergni, 1993; Uziel, 1993), participants received higher scores in grammatical sentences. Therefore, in their study for accuracy asymmetry, Bley-Vroman et al. (1988) mention that because learners tend to judge sentences as ungrammatical their accuracy rate in regard to ungrammatical sentences is raised. Again, it is not yet clear what causes accuracy asymmetry.

3. Objectives of the Study and Research Questions

The present study has intended to address the issue of the reliability of GJTs and also explores what the relationship between timed and delayed judgments is, in other words, the relationship between learners judgments based on their implicit knowledge or their explicit knowledge is another matter of concern in this study. Thus, an attempt has been made to find appropriate answers to the following research questions:

1. Do grammaticality judgment tests generate reliable scores?

2. Is there a relationship between timed and delayed judgments?
4. Method

4.1. Participants

This study was conducted with 30 EFL learners (male and female) who were selected from a population pool of 80 EFL learners majoring in English language teaching, MA level at two universities in Iran. The participants were between 24 and 37 years old.

For this study, relatively advanced EFL learners were chosen. In order to select advanced learners, an Oxford Placement Test (OPT) (Allan, 2004). was used. The reason for choosing advanced EFL learners is that in the previous studies, researchers observed the difficulty of using such tests with beginners (Ellis, 1991; Sorace, 1985). The other reason is the difficulty of the structure (verb complements) which was under focus in this study. According to Burt and Kiparsky (1972) although verb complements are taught frequently, yet all L2 learners usually have problem with these grammatical structures. Therefore, the advanced EFL learners were selected based on their performance on OPT. According to the OPT associated rating levels chart those whose scores were 75 or more were considered as advanced learners.

4.2. Instruments

4.2.1. Target Grammatical Structure

The target grammatical structure used in this study was verb complementation (i.e., infinitives, gerunds, and those clauses that function as the object of a main finite verb). Verb complements were chosen because, as it was mentioned before most of the L2 learners usually have problem with these grammatical structures. For example L2 learners usually say she admitted to tell lies, instead of she admitted telling lies.

4.2.2. OPT (Allan, 2004)

In order to select advanced EFL learners, an Oxford Placement Test was used. This test consists of 200 questions, 100 listening and 100 grammar questions but only the grammar part was used for the purpose of this study.

4.2.3. Computerized GJT

A computerized type of GJT was used in which participants were presented with 34 sentences containing verb complements. Fourteen main verbs were employed in this test (deny, suggest, admit, avoid, catch, keep, see, hope, offer, allow, decide, expect, want, encourage). All these verbs, except ‘deny’, ‘imagine’, ‘avoid’ and ‘encourage’ were selected from the first thousand frequent words based on The Teachers Word Book of 30,000 Words (Thorndike & Lorge, 1944). ‘Deny’, ‘imagine’ and ‘avoid’ were selected from second thousand words and ‘encourage’ from the third thousand words. These fourteen main verbs were used in 34 sentences which contain six verb complement types (Gerund, present participle, infinitive with object, infinitive without object, finite complement and bare infinitive). The sentences in the GJT were grouped in the following ways reflecting different complement types:

In GJT, for main verbs: ‘offer’, ‘expect’, ‘keep’, ‘catch’, ‘allow’, ‘want’, ‘avoid’ and ‘encourage’ that take only one complement type (non-finite complement) there were two sentences (one grammatical and the other ungrammatical). For these main verbs (except: ‘avoid’, ‘want’ and ‘encourage’) both the grammatical and ungrammatical sentences were grouped under the complement type that each main verb takes. Therefore, both the grammatical and ungrammatical sentences for the verb ‘offer’ were grouped as infinitive without object, for the verbs ‘expect’ and ‘allow’ as infinitive with object and for verbs ‘keep’ and ‘catch’ were grouped as present participle. In the case of main verbs ‘avoid’, ‘encourage’ and ‘want’ a non-finite complement (e.g., gerund or infinitive) was used in the grammatical sentences and a finite complement (that clause) was used in the
ungrammatical sentences. Therefore, the grammatical sentence for ‘avoid’ was grouped under *gerund* and for ‘want’ and ‘encourage’ were grouped under *infinitive with object*. Three ungrammatical sentences for these verbs were grouped under *finite complements* (e.g., that clause). For main verbs: ‘suggest’, ‘hope’, ‘decide’, ‘admit’, ‘deny’ and ‘see’, there were three sentences in the GJT (two grammatical sentences and one ungrammatical sentence). Because these verbs can take either a non-finite complement or a finite complement as their objects. For these verbs except ‘see’ one of the grammatical sentences and the ungrammatical sentence were grouped under the non-finite complement type that each main verb takes and other grammatical sentences were grouped under the finite complement type. Therefore the ungrammatical sentences and one of the grammatical sentences for verbs ‘hope’ and ‘decide’ were grouped under *infinitive without object* and for ‘suggest’, ‘admit’ and ‘deny’ under *gerund*. The other grammatical sentences for these verbs were grouped under *finite complement (that clause)*. In the case of main verb ‘see’ one of the grammatical sentences and the ungrammatical sentence were grouped under the *present participle* and the other grammatical sentence was grouped under the *bare infinitive*.

Thus, of the total of 34 sentences, 20 sentences were grammatical and 14 sentences were ungrammatical. The content of the test and the number of grammatical and ungrammatical sentences and the number of each complement type are shown in Table 1.

**Table 1. Content of the GJT**

<table>
<thead>
<tr>
<th>Type of verb comp.</th>
<th>Main verbs</th>
<th>NO. Of Gramm. S &amp; Ungramm. S</th>
<th>Total NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gerund</td>
<td>Gramm. S deny-admit-suggest-avoid</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Ungramm. S deny-admit-suggest</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Pres. Participle</td>
<td>Gramm. S Catch–keep–see</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Ungramm. S Catch–keep–see</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Inf. w/o obj.</td>
<td>Gramm. S Hope-offer–decide</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Ungramm. S Hope-offer–decide</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Inf. w/ obj.</td>
<td>Gramm. S Allow-expect-want-encourage</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Ungramm. S Allow-expect</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Finite Complement</td>
<td>Gramm. S deny-admit-suggest-hope-decide</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Ungramm. S Avoid-want-encourage</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gramm. S see</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Bare inf.</td>
<td>Ungramm. S -</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

4.3. Procedure

At first OPT was administered to a group of EFL learners (N=80) in order to select 30 advanced level learners from the MA EFL learners at two universities in Iran. Based on the OPT associated rating levels chart, those whose scores were 75 or more were considered as advanced EFL learners.

Then advanced Learners were asked to make judgments about 34 sentences in a timed GJT (timed GJT1). To this end, a computerized GJT software was designed in which learners took the test individually on a personal computer. Each sentence appeared on the monitor and stayed there for six seconds, and during this time learners made one of three possible responses, grammatical, ungrammatical, or not sure and then pressed the answer key to record their answers.

In a study by Bialystok (1979), learners were given three seconds to make immediate judgments. In present study based on the pilot study which was conducted with a group of EFL learners (N=10) with the same characteristics as those of the population who participated in the main study, six seconds was considered as sufficient time for learners to make grammaticality judgments based on their implicit knowledge (i.e., not based on their explicit knowledge) and then to press the answer key to record their answers on the computer. Before starting the test, the learners read the instruction on the screen and practiced with six example sentences.
After timed GJT1, learners were asked to correct the ungrammatical sentences in order to see to what extent their judgments were related to the structure (verb complements) under investigation. Two weeks later they took the same timed GJT (timed GJT2) on the computer again in order to examine test-retest reliability of timed GJTs.

Two weeks after the administration of the timed GJT2, each learner was asked to do the same GJT on computer again but this time there was no time limit. Each sentence appeared and stayed on the monitor until the participants selected one of the responses, then the next sentence appeared. This time participants were given as much time as they needed to make their judgments on the delayed GJT.

5. Results

5.1. Reliability of the Timed GJTs

In order to answer the first research question, the reliability of the GJTs was examined in three ways: (1) Test-retest reliability was run by comparing scores obtained from timed GJT1 with scores from timed GJT2. (2) Internal consistency reliability was measured using Cronbach's coefficient alpha for timed GJT1, timed GJT2 and delayed GJT. (3) Reliability was also examined through an analysis of the response patterns in the timed GJT1 and GJT2.

5.1.1. Test-retest Reliability of Timed GJTs

Test-retest reliability was estimated by comparing scores obtained from timed GJT1 with scores from timed GJT2. For this purpose, test-retest reliability was examined by reporting a measure of test-retest reliability for the test as a whole, then for each of the complement types and finally for the individual sentences in the two tests.

First, Pearson product–moment correlations were computed between the mean scores obtained from the two administrations of the timed GJTs by 30 participants. The overall correlation between the two test scores was $r(30) = .399, p<0.05$. This result indicates that the overall test-retest reliability of the timed GJTs was low (Garrett, 1965). The learners seem not to be consistent between the two administrations of the same GJT.

Second, the test-retest reliability coefficients for each of the complement types were computed in order to examine whether each of the complement types varies in terms of their reliabilities. The mean scores of each complement type for each participant (N=30) in the two timed GJTs were used in these analyses. (The complement type bare infinitive was excluded from the analysis because there was only one sentence for this complement in the test.) The result of test-retest reliability coefficients for five complement types revealed that there were significant but low correlations only for gerund, infinitive without object and that clause complements. The correlation coefficient for gerund, infinitive without object and that clause complements were $r (30) = .367, p<0.05; r (30) = .469, p<0.05; r (30) = .461, p<0.05$ respectively (only the reliabilities for infinitive without object and finite complement (that clause) exceed 0.4, which is considered not to be acceptable according to Garrett (1965)). Therefore, the test-retest reliabilities for each of the complement types were also low.

Third, test-retest reliabilities were examined for the individual sentences in the timed GJT1 and GJT2. To this end, the participants’ correct and incorrect responses on the two tests were compared and the correlations were computed. For this analysis, not sure and late responses (i.e., the responses not made within the 6 seconds) were considered as incorrect responses. The reason for such decision is that both not sure and late responses indicated lack of grammatical knowledge. Table 2 presents the test-retest reliability coefficients for individual sentences between GJT1 and GJT2 (only sentences with significant correlations are included in the table.)

<table>
<thead>
<tr>
<th>Item</th>
<th>GJT2-1</th>
<th>Item</th>
<th>GJT2-1</th>
<th>Item</th>
<th>GJT2-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>9) GJT1-1</td>
<td>r</td>
<td>.592**</td>
<td>13) GJT1-1</td>
<td>r</td>
<td>.396*</td>
</tr>
<tr>
<td>25) GJT1-1</td>
<td>r</td>
<td>.590*</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As table 2 shows, there were only 10 statistically significant correlations out of 34 correlations (sentences 9, 10, 11, 13, 15, 18, 25, 27, 31, and 32). However, the reliability coefficients for these sentences were also at the lower end of being acceptable (Garrett, 1965). In other words, the test-retest reliabilities for individual sentences indicated a low level of reliability for the test in the timed GJT1 and GJT 2.

5.1.2. Internal Consistency Reliability of the Timed GJTs

Internal consistency reliability was measured using Cronbach's coefficient alpha for timed GJT1, timed GJT2 and delayed GJT. The computed Cronbach's coefficient alpha for timed GJT1, timed GJT2 and delayed GJT were .593, .382, .440 respectively. Therefore, the internal consistency reliability for timed GJT1, timed GJT2 and delayed GJT was also low (The internal consistency reliability only for timed GJT1 and delayed GJT exceeded 0.4, however, they were also at the low side of being acceptable according to Mehrens and Lehmann (1973)).

5.1.3. Response Patterns in the Timed GJTs

To examine the response patterns in the timed GJTs, First, the numbers of grammatical, ungrammatical, not sure and late responses in timed GJT1 and 2 were counted (The late responses were included in the not sure responses). Table 3 shows the frequencies of the different types of response in timed GJT1 and 2.

<table>
<thead>
<tr>
<th></th>
<th>grammatical</th>
<th>ungrammatical</th>
<th>Not sure/ late responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timed GJT 1</td>
<td>514</td>
<td>399</td>
<td>107 (27,80)</td>
</tr>
<tr>
<td>Timed GJT 2</td>
<td>545</td>
<td>399</td>
<td>76  (15,61)</td>
</tr>
</tbody>
</table>

As this table displays, in timed GJT1, the numbers of grammatical, ungrammatical and not sure responses were 514, 399, and 107 respectively. In timed GJT2, the participants responded grammatical for 545 sentences, ungrammatical for 399 sentences and not sure for 76 sentences. The participants provided not sure responses only 2.65% of the time in timed GJT1 and 1.47% of the time in timed GJT2. And the participants made late responses for 80 sentences in timed GJT1 and 61 sentences in timed GJT2. These results revealed that although participants were uncertain of their judgments, they were reluctant to choose not sure response.

Second, in order to examine whether the participants were biased toward grammatical or ungrammatical sentences, a Chi-square test was run. The not sure responses were excluded from this analysis based on this assumption that a not sure response was not either a grammatical or an ungrammatical response. The results of the Chi-Square test for participants’ tendency toward grammatical or ungrammatical responses showed that, the Chi-square statistic (=.334) was not significant at the 0.05 level (P=.563). In other words, the differences between observed and expected frequencies were not significant. Thus, it can be concluded that the participants were not biased toward accepting the sentences as grammatical or ungrammatical in this study.

Next, the participants’ performance on the grammatical and ungrammatical sentences in all three tests was examined in order to determine the accuracy of participants' judgments. The descriptive statistics for the participants’ responses to the grammatical (N=20) and ungrammatical (N=14) sentences in timed GJT1, GJT2 and
delayed GJT demonstrated that, the mean scores for grammatical sentences in all three tests (GJT1, GJT2 and delayed GJT) were higher than mean scores for ungrammatical sentences. Therefore, the descriptive statistics suggested that participants scored higher with grammatical sentences in all three tests. In order to see whether these differences in mean scores of grammatical and ungrammatical sentences in all three tests were significant or not a MANOVA test was run. Table 4 displays the Results of MANOVA analysis for the overall item effect.

Table 4. Results of MANOVA Analysis for the Overall Item Effect

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>F</th>
<th>Hypothesis df</th>
<th>Error df</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pillai's trace</td>
<td>.692</td>
<td>41.849</td>
<td>3.000</td>
<td>56.000</td>
<td>.000</td>
<td>.692</td>
</tr>
<tr>
<td>Wilks' lambda</td>
<td>.308</td>
<td>41.849</td>
<td>3.000</td>
<td>56.000</td>
<td>.000</td>
<td>.692</td>
</tr>
<tr>
<td>Hotelling's trace</td>
<td>2.242</td>
<td>41.849</td>
<td>3.000</td>
<td>56.000</td>
<td>.000</td>
<td>.692</td>
</tr>
<tr>
<td>Roy's largest root</td>
<td>2.242</td>
<td>41.849</td>
<td>3.000</td>
<td>56.000</td>
<td>.000</td>
<td>.692</td>
</tr>
</tbody>
</table>

As table 4 displays, the MANOVA test showed an overall item (grammatical vs. ungrammatical) effect in the three GJTs. The multivariate F ratio (=41.849) for the overall item effect was significant (p<0.05) and the effect size (partial eta squared = .692) was also large (based on the guidelines proposed by Cohen (1988) for interpreting the eta squared value, 0.01 = small effect, 0.06 = moderate effect, 0.14 = large effect). Therefore, from MANOVA analysis, it can be concluded that the differences in mean scores of grammatical and ungrammatical sentences in all three tests were significant, the participants scored significantly higher with the grammatical sentences in all three tests.

All univariate F ratios for the item effect in the three GJTs were also examined. The results of univariate analysis confirmed the results of multivariate analysis. All univariate F ratios for the item effect in the three GJTs (F=52.382, P=.000; F=92.724, P=.000; F=71.297, P=.000) were also significant. And their effect sizes (partial eta squared = .475, .615, .551) were also large. It means that learners scored significantly higher with the grammatical sentences in all three tests. Therefore, the inferential statistics also confirmed the descriptive finding mentioned previously.

Finally, the responses that changed from timed GJT1 to GJT2 were computed in order to examine the nature of the changed responses in the two administrations of the timed GJTs. The learners changed 360 responses from timed GJT1 to GJT2. This was 35.29% of the total responses. A Pearson product-moment correlation was computed between the scores of each learner and the number of changes which the learner made. There was a significant negative correlation (r (30) = -.756, p<.000) between the scores of learners and the number of their changes. It means that the scores of the learners changed significantly with the number of changes they made in their judgments. Their scores were higher if they had more stable rules and did not make many changes in their judgments. And their scores were lower if they made more changes in their judgments. The more proficient learners had more stable rules and made more consistent judgments.

5.2. Relationship between Timed and Delayed Judgments

In order to answer the second research question the relationship between timed and delayed judgments was examined. First, the mean scores obtained on timed GJT1 and the delayed GJT by the 30 participants were correlated using the Pearson product–moment coefficient. There was a low significant correlation (r (30) = .470, p<0.05) between the two test scores. This means that there was a degree of relationship between timed GJT1 and the delayed GJT; however, this relationship was not strong.

Next, correlations were examined for the individual sentences in timed GJT1 and the delayed GJT. The reason for this analysis is that a better way to investigate the relationship between timed GJT1 and the delayed GJT is looking at the correlations involving individual sentences. Table 5 presents the result of the Pearson product–moment coefficients for individual sentences in timed GJT1 and the delayed GJT. (Only sentences with significant correlations are included in the table 5)
Table 5. The Pearson product–moment Coefficients for the Individual Sentences in Timed GJT 1 and the Delayed GJT

<table>
<thead>
<tr>
<th>Item</th>
<th>GJTD-1</th>
<th>Item</th>
<th>GJTD-1</th>
<th>Item</th>
<th>GJTD-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>8) GJT1-1</td>
<td>r .428*</td>
<td>18) GJT1-1</td>
<td>r .373*</td>
<td>25) GJT1-1</td>
<td>r .436*</td>
</tr>
<tr>
<td>p .018</td>
<td>p .042</td>
<td>p .010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9) GJT1-1</td>
<td>r .671**</td>
<td>21) GJT1-1</td>
<td>r .380*</td>
<td>26) GJT1-1</td>
<td>r .443*</td>
</tr>
<tr>
<td>p .000</td>
<td>p .038</td>
<td>p .012</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24) GJT1-1</td>
<td>r .438*</td>
<td>29) GJT1-1</td>
<td>r .671*</td>
<td>30) GJT1-1</td>
<td>p .000</td>
</tr>
<tr>
<td>p .015</td>
<td>p .012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As this table shows, there were only 8 statistically significant correlations out of 34 computed correlations. Thus, the relationship between the timed and delayed judgments was also weak at sentence level.

6. Discussion

The results of data analysis revealed that: (1) the overall test-retest reliability of the timed GJTs was low; the learners were only moderately consistent between two administrations of the same timed GJT. (2) The test-retest reliabilities for each complement type were low. (3) The test-retest reliabilities of individual sentences in the timed GJTs were fairly low and this indicated low level of reliability for the test. (4) The internal consistency reliability for timed GJT1, timed GJT2 and delayed GJT was also low. In general, these results showed that the GJT used in this study lacks high level of reliability. These results are not in line with those of Mandell (1999). In his study; GJTs were compared with dehydrated sentences test (DST) (an assessment tool commonly used in the L2 classrooms). Data were collected from adult L2 learners of Spanish about verb movement, the result from the comparison of the two tests revealed that GJTs were reliable measures of linguistic knowledge. Regarding the response patterns in the timed GJTs the following results were observed in this study: (1) The learners did not demonstrate a response bias toward grammatical or ungrammatical responses. This result is in contrast with those of Bley-vorman et al. (1988), They noted that their participants judged sentences as ungrammatical more often than as grammatical regardless of their grammaticality. (2) In timed GJT1 and 2 the learners scored higher with grammatical sentences than with ungrammatical ones. It seems that learners had greater difficulty with the ungrammatical sentences than with the grammatical ones. This result is in accordance with those of the previous studies about accuracy asymmetry in judging grammatical vs. ungrammatical sentences (Ellis, 1991; Hawkins et al, 1993; Uziel, 1993). They also showed that their learners scored higher with grammatical sentences than with ungrammatical ones. (3) The learners were reluctant to choose the not sure response. Therefore, their definite judgments (grammatical or ungrammatical) may be based on their grammatical knowledge or just a strategy to avoid using the not sure response. In terms of the relationship between timed GJT1 and the delayed GJT, (1) the overall relationship between timed GJT1 and the delayed GJT was significant but the correlation was not a strong indication of connection between the two tests. (2) The investigation of the correlations involving individual sentences on the two tests revealed that the relationship between timed GJT1 and the delayed GJT at the sentence level was also weak. These results are in line with those of Han (2000), who concluded that the relationship between timed GJT and untimed GJT was a weak one. One possible explanation for this result may be that learners may have used different types of knowledge in the two tests.

7. Conclusion

Constructing a reliable test has always been a goal for test designers. According to Wells and Wollack (2003) “Test reliability refers to the consistency of scores students would receive on alternate forms of the same test”, (p.2). The results of this study indicated that the GJT used in this study was not reliable enough to be used with confidence in order to measure EFL learners’ grammatical competence. The low level of test-retest reliability of the timed GJTs and also the low level of internal consistency reliability of all three GJTs, suggest that researchers, teachers and those who are interested in the use of GJTs, should use such tests with more attention and caution. Analysis of response patterns also revealed that participants were not consistent in their judgments, all the
participants made a considerable number of changes (35.29%) from GJT1 to GJT2. And also they were reluctant to select the not sure response. Even when they were uncertain they preferred to guess rather than to choose not sure response. These results suggest that GJT data are not an exact reflection of learners’ grammatical knowledge, as it was believed before (e.g., Carroll & Meisel, 1990; Ellis, 1991). The weak relationship between timed GJT and delayed GJT suggest that participants may use different types of knowledge under different test administration conditions. In timed GJT as time is limited and participants are under pressure to make their judgments in a few seconds it seems that they rely on their intuitions and implicit knowledge rather than explicit knowledge, but in delayed GJT, since students are free to make judgments in as much time as they want, their judgments seem to reflect using more explicit rules. However, this study was undertaken with a small number of samples and only one type of GJT. Therefore, similar studies with more samples and other types of GJTs can be recommended.

References